$\mathcal{S}$=*fl*(approx(S)) & 0.0000e00≠0

Forward-Backward Heat Equations and Analysis of Iterative Methods

Hao Lu

Forward-Backward Heat Equations and Analysis of Iterative Methods

een wetenschappelijke proeve op het gebied van de
Wiskunde en Informatica

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen,
volgens besluit van het College van Decanen
in het openbaar te verdedigen
op dinsdag 10 Oktober 1995
des namiddags te 15.30 uur precies

door

Hao Lu

geboren op 1 November 1961 te Shaanxi, Volksrepubliek China

Promotor: Prof. Dr. A.O.H. Axelsson

Het onderzoek dat tot dit proefschrift heeft geleid, werd gesteund door de Nederlandse Stichting voor de Wiskunde SMC met een subsidie van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

CIP-GEGEVENS KONINKLIJKE BIBLIOTEEK, DEN HAAG

Lu, Hao

Forward-Backward Heat Equations and Analysis of Iterative Methods / Hao Lu. - [S.l. : s.n.]. - Ill.
Thesis Katholieke Universiteit Nijmegen. - With ref.
ISBN 90-9008653-6
Subject headings: numerical analysis

To my wife, Jian

Contents

Summary

The thesis contains two parts including the following eight papers (eight chapters) finished during the first two and a half years from the end of 1992 to the beginning of 1995 of my stay in the Netherlands as a Ph.d student under the guidance of Professor O. Axelsson.

[LW] (with Z.-Y. Wen) *Solution of a forward-backward heat equation*, (submitted)

[L1] *Galerkin and weighted Galerkin methods for a forward-backward heat equation*, (submitted)

[LM] (with J. Maubach) *A finite element method and variable transformations for a forward-backward heat equation*, (manuscript).

[L2] *A uniform-consistency barrier on finite-difference schemes of positive type for convection-diffusion equations*, SIAM J. Sci. Comput., 16 (1995), 169–172.

[ALP] (with O. Axelsson and B. Polman) *On the numerical radius of matrices and its application to iterative methods*, Lincar and Multilinear Algebra, 37 (1994), 225–238.

[L3] *Matrix compensation and diagonal compensation*, J. Comp. Appl. Math., (to appear).

[AL] (with O. Axelsson) *On eigenvalue estimates for block incomplete factorization methods*, SIAM J. Matrix Anal. Appl. 16 (1995), (to appear).

[LA] (with O. Axelsson) *Conditioning analysis of incomplete block factorizations and its application to elliptic equations*, (submitted)

Part 1 deals with the existence and uniqueness of a weak solution, finite element methods for forward-backward heat equations and a uniform-consistency barrier on finite-difference schemes of positive type for convection-diffusion problems consisting of the first four chapters (papers [LW], [L1], [LM] and [L2]). Part 2 which includes the remaining chapters deals mainly with the use of the numerical radius to analyze the rate of convergence of iterative methods for nonsymmetric linear systems, generalization of diagonal compensation, estimates of condition number and eigenvalues for block incomplete factorization methods and their applications to elliptic equations. I summarize the thesis part by part.

1. Forward-Backward Heat Equations and a Barrier of Finite Difference Schemes

Many finite element methods for heat equations have been proposed and analyzed in the literature (cf Thomée [27]). A common approach for a heat equation is first to apply a Galerkin method in space to reduce the equation to a set of ordinary differential equations. Then a suitable method is applied to integrate the ordinary differential equations. Unfortunately, some heat equations do not fit into

the standard methods, for example, forward-backward heat equations

$$\sigma u_t = r u_{xx} - q u_x - p u + f, \quad \forall (x,t) \in \Omega, \tag{0.0.1}$$

$$\begin{cases} u(\mu_1(t),t) = u(\mu_2(t),t) = 0, & \forall t \in (0,T), \\ u(x,0) = u_0(x) & \text{for } \sigma(x,0) > 0, \\ u(x,T) = u_T(x) & \text{for } \sigma(x,T) < 0, \end{cases} \tag{0.0.2}$$

where $\Omega = \{(x,t) : \mu_1(t) < x < \mu_2(t),\ 0 < t < T\}$, $\mu_1(t)$ and $\mu_2(t)$ are continuous and piecewise smooth functions, σ, r, q, p, u_0, u_T are known functions, $r \geq \rho > 0$. The diffusion coefficient σ changes sign in Ω.

Problem (0.0.1), (0.0.2) arises in various applications, such as, boundary layer problems in fluid dynamics [25], [26], plasma physics and astrophysics in the study of propagation of an electron beam through the solar corona [19].

Problems like $\sigma(x,t)u_t - u_{xx} = f(x,t)$ with $\sigma(x,t)$ taking different signs were first considered by Gevrey [13], [14]. He treated in particular the case $\sigma(x,t) = x^m$ with m an odd integer. Later, in 1968, Baouendi and Grisvard [10] dealt with the case $\sigma(x,t) = x$ in detail. A similar treatment in a context where the second-order derivative is replaced by a suitable nonlinear differential operator can be found in Lions' book [20]. Franklin and Rodemich [12] considered also the case $\sigma(x,t) = x$ and treated the equation on $-\infty < x < \infty$, $0 < t < T$. In 1976, Pagani showed the existence and uniqueness of a weak solution of problem (0.0.1), (0.0.2) for a special case $\sigma = \omega(x)$, $\omega(x)\operatorname{sgn} x > 0$ for $x \neq 0$, $\mu_1(t) = a$, $\mu_2(t) = b$ and $u_0 = u_T = 0$. In 1984, Goldstein and Mazumdar [15] gave also an existence and uniqueness theorem of a weak solution for the equation (0.0.1) subject to

$$\begin{cases} u(a,t) = u(b,t) = 0, & \forall t \in (0,T), \\ u(x,0) = u_0(x), & \forall x \in (0,b), \\ u(x,T) = u_T(x), & \forall x \in (a,0), \end{cases} \tag{0.0.3}$$

under certain assumptions. Unfortunately, the problem (0.0.1), (0.0.3) is overdetermined in general even if their assumptions are satisfied.

To understand forward-backward heat equations correctly, our task in Chapter 1 is to show the existence and uniqueness of a weak solution for the problem (0.0.1), (0.0.2) in a certain Hilbert space and show that the problem is well-posed in a certain sense. The proof is based a variant of the Lax-Milgram lemma due to Lions [20]. Comparisons with previous result, i.e., Pagani's result [24], on the existence and uniqueness of solution of (0.0.1), (0.0.2), the results presented in Chapter 1 have the following advantages:

- The new result is applicable to a much wider class of equation than Pagani's result. The assumption made on the coefficients of equation is weaker than Pagani's assumption even if $\sigma = \omega(x)$, $\omega(x)\operatorname{sgn} x > 0$ for $x \neq 0$ and $u_0 = u_T = 0$. The new result and Pagani's result on the existence and uniqueness are given in different sense, but the new one is much stronger than the old one. Pagani's result follows immediately from the result in Chapter 1.
- The new result provides a mathematical basis for finite element methods for problem (0.0.1), (0.0.2) in Chapters 2 and 3, but Pagani's result cannot be used for the same purpose.

In the past few years, some authors have already paid attention to numerical solution methods for the model problem of $\mu_1(t) = -1$, $\mu_2(t) = 1$ and $p = q = u_0 = u_T = 0$. In 1990, Vanaja and Kellogg [29] presented some iterative methods to solve finite difference approximation to (0.0.1), (0.0.2) if $\sigma(x,t) = \sigma(x)$ or $\sigma_t(x,t) \leq 0$. In [8], [9] and [21] Aziz and Liu considered some other numerical methods for (0.0.1), (0.0.3). Though problem (0.0.1), (0.0.3) may have no solution, Aziz and Liu's numerical methods can be used to solve (0.0.1) and (0.0.2) with some straightforward modifications. Their approaches are to transform the problem to a first-order system of partial differential equations and to solve the systems in $(L^2(\Omega))^2$ by a Galerkin method [8] and a least squares method [9].

Following the analysis of the existence and uniqueness we focus our attention on finite element methods, namely space-time methods, for forward-backward heat equations. In Chapter 2, we present Galerkin and weighted Galerkin methods for (0.0.1), (0.0.2) without transforming the equation to a first-order system of partial differential equations for the model case, i.e., $\mu_1(t) = -1$, $\mu_1 = 1$ and $p = q = u_0 = u_T = 0$. We consider a simultaneous discretization of space and time variables by using continuous finite element methods. If there exist two functions g, $q \in H^1(\Omega)$ such that $|q| \leq C < +\infty$ and

$$\frac{1}{2}g_x - \frac{1}{2}\sigma_t - \sigma q_t - q_x^2 \geq \frac{\beta}{4}g^2, \tag{0.0.4}$$

which holds in particular for $\sigma_t(x,t) \leq b < \pi^2/2$, our results show that the L^2 rate of convergence is $O(h^k)$, where h is the meshsize of space and time, if the solution $u \in H^{k+1}(\Omega)$ and piecewise polynomials of degree k are used. The linear systems of the discrete equations arising from the methods are positive definite. Comparison with previous methods known for the forward-backward heat equations, for example Vanaja-Kellogg's method [29] and Aziz-Liu's method [8]. The methods presented in this thesis have the following advantages:

- The new methods can be used to solve a much wider class of equations than Vanaja-Kellogg's method and Aziz-Liu's methods. The assumption made on the coefficients of the equation is weaker than previous ones. Aziz-Liu's assumptions are stronger than (0.0.4) if their method is used to solve (0.0.1), (0.0.2). Their assumptions imply the inequality (0.0.4). Furthermore, it is shown that if Aziz and Liu's method is applicable to (0.0.1), (0.0.2), so is the weighted Galerkin method. The difference between $\sigma_t < \pi^2/2$ and Vanaja-Kellogg's assumption $\sigma_t \leq 0$ is essential. A example shows that doing a transformation $y = y(t)$ for a wide class of equations (0.0.1), (0.0.2) with $\sigma_t \geq \pi^2/2$ for some points in Ω, we obtain new forward-backward heat equations (0.0.1), (0.0.2) such that the corresponding σ satisfies $\sigma_t < \pi^2/2$, but there is no transformation $y = y(t)$ such that the corresponding σ satisfies $\sigma_t \leq 0$.
- The new methods require less regularity for the solution of the equations than the previous numerical methods. Vanaja-Kellogg's method requires that the solution possesses a continuous derivative of order 4 in x and order 2 in t to obtain the rate of convergence $O(k+h^2)$, where k and h are meshsize in time and in space, respectively. Aziz-Liu's method requires the solution $u_t, u_{tx}, u_{xx} \in L^2(\Omega)$ if it is used to solve (0.0.1), (0.0.2). The new methods need only the solution $u \in H^1(\Omega)$.

- Both Galerkin and weighted Galerkin methods need fewer computations than Aziz-Liu's method because they only involve half the number of unknowns compared with Aziz-Liu's method [8].
- Unlike Aziz-Liu's methods, the new methods do not need to preprocess the boundary condition to match the boundary condition required by the corresponding first-order systems. It still remains unknown how to do such preprocessing.

In Chapter 3, we generalize the methods to the general case. The generalization is based on the result of the existence and uniqueness of a weak solution for (0.0.1), (0.0.2) given in Chapter 1 and possesses all advantages mentioned above. To solve a wide class of the equations (0.0.1), (0.0.2), we do variable transformations $x = x$ and $y = y(t)$ for the equation such that the new equation can be solved by the Galerkin method. We derive some conditions under which we can do the transformation to solve a wide class of equations (0.0.1), (0.0.2) and show how to construct the transformations. In particular, the conditions are automatically satisfied if σ is separable, i.e., $\sigma = \kappa(x)\varphi(t)$. For other important cases where $\sigma(x,t)$ is a function of $x+ct+d$, i.e., $\sigma(x,t) = \sigma(x+ct+d)$ we first do variable transformations $y = x+ct+d$ and $t = t$. Then using a simple function transformation we obtain a forward-backward heat equation which satisfies all assumptions for our generalization.

It is well-known that finite difference schemes of positive type are efficient for convection-diffusion problems. Unfortunately, in 1978, Kellogg and Tsan showed that any 3-point difference schemes of positive type cannot approximate $L(u) = -\varepsilon u'' + b(x)u' + g(x)u$ to $O(h^\alpha)$ $(\alpha > 1)$ accuracy uniformly in ε [18]. Note that this barrier can be overcome somehow in one dimension by slightly adding meshpoints depending on ε. Recently, Axelsson and Nikolova showed an $O(h^2)$ accuracy uniformly in ε by using $O(\log \varepsilon^{-1} + h^{-1})$ points [7]. In the last chapter of this part we show a uniform-consistency barrier on finite difference schemes of positive type for convection-diffusion equations, i.e., any difference scheme of positive type cannot approximate $Lu = -\varepsilon \Delta u + \vec{f} \cdot \nabla u + gu$ to $O(h^\alpha)$ $(\alpha > 1)$ accuracy uniformly in ε.

2. Analysis of Iterative Methods.

The concern in this part is to analyze the basic iterative methods and SOR method for quasi-Hermitian positive definite matrices by using numerical radius, to generalize diagonal compensation for symmetric positive definite matrices to positive definite matrices and to estimate eigenvalues of block incomplete factorization methods and compensative preconditioners.

Commonly the spectral radius is used to analyze the rate of convergence of iterative methods. However, for a nonsymmetric iteration matrix B this gives only information about the asymptotic rate of convergence. In Chapter 5, it is shown first that the numerical radius $r(B)$ is a more reliable measure of the convergence behavior for the initial iterations, because

$$r(B^m)^{\frac{1}{m}} \le \|B^m\|^{\frac{1}{m}} \le 2^{\frac{1}{m}} r(B^m)^{\frac{1}{m}} \le 2^{\frac{1}{m}} r(B).$$

A matrix A is called a quasi-Hermitian positive definite if there exists a nonsingular block diagonal matrix P such that PAP^{-1} is Hermitian positive definite. Next in the analysis of the successive overrelaxation method for quasi-Hermitian positive definite matrices it is shown that a crucial parameter (γ) depends on the numerical radius of the (block) lower triangular part $(\tilde{L})$ of the standard splitting of the matrix.

It is shown also, using the numerical radius, that one can derive an upper bound of the largest eigenvalue of the preconditioned matrix $C^{-1}A$, when A is a symmetric and positive definite matrix partitioned in $m \times m$ blocks and C is an block incomplete factorization of A. Under a certain condition this upper bound is $2m$.

When we construct a preconditioner for a positive definite matrix, it is important to preserve the positivity of matrices. We introduce matrix compensation for the purpose. Let A be a positive definite matrix. Roughly speaking, matrix compensation is to find a matrix G such that $G - A$ is positive semidefinite. In general, it is not easy to find a good compensative matrix G to construct a good preconditioner for A by using G. On some occasions, however, the diagonal compensation can be easily used to obtain matrix compensation for symmetric positive definite matrices for constructing good preconditioners [4], [5]. It is well known that diagonal compensation acts as a key in modified block incomplete factorization. We show how to do diagonal compensation for nonsymmetric matrices.

Let $A = B + R$ be a symmetric positive definite matrix and $M = B + D$, where B is a symmetric matrix and D is a diagonally compensative matrix of R. M can be an efficient preconditioner of A (see, e.g., [4], [5]). If $B\mathbf{v} > 0$ for some positive vector $\mathbf{v}$ or $\rho(A^{-1}R) < 1$, some general results on upper bounds of eigenvalues and condition number for $M^{-1}A$ are derived by using the spectral radius $\rho(B^{-1}R)$ and the condition number $\kappa(M^{-1}B)$ in [4]. Sometimes, however, it is difficult to estimate $\rho(B^{-1}R)$ and $\kappa(M^{-1}B)$ accurately. In Chapter 6, we show some new upper bounds of eigenvalues and condition numbers for compensative preconditioners involving $\rho(B^{-1}D)$ or $\rho(M^{-1}D)$. Condition numbers of preconditioned matrices are estimated if the matrix compensation is used to construct preconditioners for symmetric positive definite matrices.

To estimate the rate of convergence of preconditioned iterative methods such as the Chebyshev iterative method and the conjugate gradient iterative method, one needs to know the extreme eigenvalues and the distribution of eigenvalues of the preconditioned matrix respectively, see [1], [2], [17], [16], [6], [28]. Naturally, this problem by itself is difficult, especially for the distribution of all eigenvalues. Fortunately, it has been shown (see [3]) that under certain conditions lower and upper bounds of some eigenvalues can be derived and they provide the information necessary to compare modified and unmodified incomplete factorization methods for symmetric positive definite matrices, for instance.

Consider the implicit preconditioner on factorized form

$$C = (X + L)X^{-1}(X + L^T)$$

of a symmetric matrix A. Let $A = D_A + L_A + L_A^T$, where D_A is a block diagonal matrix. If A is a Stieltjes matrix and $L = L_A$ in some cases, some methods to estimate upper bounds of eigenvalues of $C^{-1}A$ were derived in [11], [3], [23], [22]. However, the assumptions limit the applicability of the results because for incomplete factorization methods they do not hold in general. In Chapter 7, we discuss upper bounds and distribution of eigenvalues of block incomplete preconditioners for the general case of A being only a symmetric matrix. All of results allow that L_A differs from L. Even when the assumption of A is weakened, we can have strong results. The results here unify some of the previous results on upper bounds of eigenvalues of incomplete preconditioners. We also generalize the well-known

inequality that the spectral radius is bounded by the trace for symmetric positive semidefinite matrices to block form.

Let $\lambda_i(A)$ denote the ith eigenvalue of A and assume that the eigenvalues of a matrix are ordered in a non-increasing order. Chapter 8 continues our analysis of block incomplete factorization methods. We show that if there exist two constants α_i and σ_i such that $\alpha_i \leq \lambda_i(X^{-1}K) \leq \sigma_i$, where $K = A - L - L^T$, then

$$\lambda_{\min}(M(\alpha_i)) \leq \lambda_i(C^{-1}A) \leq \lambda_{\max}(M(\sigma_i)),$$

where $M(\alpha) = (I+\tilde{L})^{-1}+(I+\tilde{L}^T)^{-1}+(\alpha-2)(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1}$, $\tilde{L} = X^{-\frac{1}{2}}LX^{-\frac{1}{2}}$. The result yields immediately a simple upper bound $1/(2-\sigma_i)$ for the ith eigenvalue of the preconditioned matrix $C^{-1}A$ if $\sigma_i < 2$, which result generalizes the old one for the maximum eigenvalue in Chapter 7. Considering the relation between A and C and using the generalization of the well-known inequality that the spectral radius is bounded by the trace for symmetric positive semidefinite matrices to block form in Chapter 7, we show also that $m+1$ is another upper bound for the maximum eigenvalue if $\sigma_1 \leq 2$ and $A - C$ is positive semidefinite. The results are used to estimate bounds of each eigenvalue if the generalized SSOR method is used to solve an elliptic equation

$$-\frac{\partial}{\partial x}\left(a_1(x,y)\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(a_2(x,y)\frac{\partial u}{\partial y}\right) = f(x,y) \quad \text{on } \Omega, \tag{0.0.5}$$

$$u(x,y) = g(x,y) \ \text{ on } \Gamma = \partial\Omega,$$

where $\Omega = (0,a) \times (0,b)$, $a_1(x,y)$ and $a_2(x,y)$ are positive functions. By doing a transformation, we show how the coefficients $a_1(x,y)$ and $a_2(x,y)$ influence bounds of eigenvalues of the preconditioned matrix. The analysis of eigenvalue estimates yields an $O(h^{-1})$ upper bound for the condition number of the preconditioned matrix if the modified block factorization method is used for elliptic equations (0.0.5) with variable coefficients under the assumption that $\log a_1(x,y)$ satisfies the Lipschitz condition for x which implies that we allow that $a_1(x,y)$ has jumps in y-direction.

Acknowledgments

First of all I thank my supervisor Professor Owe Axelsson for his guidances and introducing me forward-backward heat equations and preconditioned iterative methods.

During the three years of my stay in Nijmegen as a Ph.d student I have been helped and influenced by many people. Gene Golub and Nicholas J. Higham offered me useful suggestions and encouragement. Nicholas J. Higham helped me in writing for the mathematics in English. William Layton suggested me the problem discussed in Chapter 4 and offered me valuable comments. Maya Neytcheva and Ben Polman helped me a lot on computers. They also offered me a lot of other helps. Ben Polman translated a short version of the summary and my curriculum vitae at the end of the thesis. I thank all these people together with many others who helped me in my mathematical career.

I thank especially my wife Jian Zhang for her assistance and support.

Finally I acknowledge the Netherlands Organization for Scientific Research (NWO) for the financial support.

References

[1] O. AXELSSON, *A class of iterative methods for finite element equations*, Comput. Methods Appl Mech. Engrg., 9 (1976), pp. 123–137.

[2] ———, *Solution of linear systems of equations: iterative methods*, in Sparse Matrix Techniques, Lecture Notes in Mathematics 572, V. A. Barker, ed., Springer Verlag, Berlin, Heidelberg, New York, 1977, pp. 1–50.

[3] ———, *Bounds of eigenvalues of preconditioned matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 847–862.

[4] ———, *Iterative Solution Methods*, Cambridge University Press, New York, 1994.

[5] O. AXELSSON AND V. BARKER, *Finite Element Solution of Boundary Value Problems*, Academic Press, Orlando, Fl., 1984.

[6] O. AXELSSON AND G. LINDSKOG, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math , 48 (1986), pp. 499–523.

[7] O. AXELSSON AND M. NIKOLOVA, *Adaptive refinement for convection-diffusion problems based on defect-correction technique and finite difference methods*, in progress, (1995).

[8] A. K AZIZ AND J.-L. LIU, *A Galerkin method for the forward-backward heat equation*, Math. Comp., 56 (1991), pp. 35–44.

[9] ———, *A weighted least squares method for the backward-forward heat equation*, SIAM J. Numer. Anal , 28 (1991), pp. 156–167.

[10] M. S. BAOUENDI AND P. GRISVARD, *Sur une équation d'évolution changeant de type*, J. Funct. anal., 2 (1968), pp. 352–367.

[11] R. BEAUWENS, *Upper eigenvalue bounds for pencils of matrices*, Linear Algebra Appl., 62 (1984), pp. 87–104.

[12] J. A. FRANKLIN AND E. R. RODEMICH, *Numerical analysis of an elliptic-parabolic partial differential equation*, SIAM J. Numer. Anal., 5 (1968), pp. 680–716.

[13] M. GEVREY, *Sur les équations aux dérivées partielles du type parabolique*, J. Math. pures Appl., 6 (1913), pp 305–475.

[14] ———, *Sur les équations aux dérivées partielles du type parabolique (suite)*, J. Math. pures Appl., 6 (1914), pp. 105–148.

[15] J. A. GOLDSTEIN AND T. MAZUMDAR, *A heat equation in which the diffusion coefficient changes sign*, J. Math. Anal. Appl., 103 (1984), pp. 355–364.

[16] A. GREENBAUM, *Comparison of splittings used with the conjugate gradient algorithm*, Numer. Math., 33 (1979), pp. 181–194.

[17] A JENNINGS, *Influence of the eigenvalues spectrum of the convergence rate of the conjugate gradient method*, IMA Journal of Numerical Analysis, 20 (1977), pp. 61–72.

[18] R. B. KELLOGG AND A TSAN, *Analysis of some difference approximations for a singular perturbation problem without turning points*, Math Comp., 32 (1978), pp. 1025–1039.

[19] T. LAROSA, *The propagation of an electron beam through the solar corona*, PhD thesis, Department of Physics and Astronomy, University of Maryland, 1986.

[20] J. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod Gauthier-Villars, Paris, 1969.

[21] J -L. LIU, *A finite difference method for symmetric positive differential equations*, Math. Comp., 62 (1994), pp. 105–118.

[22] M. M. MAGOLU, *Analytical bounds for block approximate factorization methods*, Linear Algebra Appl., 179 (1993), pp. 33–57.

[23] M. M. MAGOLU AND Y. NOTAY, *On the conditioning analysis of block approximate factorization methods*, Linear Algebra Appl., 154-156 (1991), pp. 583–599.

[24] C. M. PAGANI, *On forward-backward parabolic equations in bounded domains*, Bolletino U. M. I., (5) 13-B (1976), pp. 336–354.

[25] K STEWARTSON, *Multistructural boundary layers on flat plates and related bodies*, Adv. in Appl. Mech., 14 (1974), pp. 145–239.

[26] ———, *D'Alembert's paradox*, SIAM Review, 23 (1981), pp. 308–343.

[27] V THOMÉE, *Galerkin finite element methods for parabolic problems*, Lecture Notes in Math., vol. 1054, Springer-Verlag, 1972.

[28] A. VAN DER SLUIS AND H. A. VAN DER VORST, *The rate of convergence conjugate gradients*, Numer. Math., 14 (1986), pp. 543–560.

[29] V. VANAJA AND R. B KELLOGG, *Iterative methods for a forward-backward heat equation*, SIAM J. Numer. Anal., 27 (1990), pp. 622–635.

Part 1

Forward-Backward Heat Equations and a Barrier on Finite-Difference Schemes

CHAPTER 1

Solution of a Forward-Backward Heat Equation*

Abstract. The existence and uniqueness of a weak solution for a forward-backward heat equation are shown that provides a mathematical basis for a finite element method for the equation. The stability criterion presented in the paper shows that the problem is well-posed on a certain Hilbert space.

Key words. existence and uniqueness, forward-backward heat equation, well-posed problem

AMS subject classifications. 35K05

1.1. Introduction

Consider a forward-backward heat equation

$$\sigma u_t = r u_{xx} - q u_x - p u + f, \quad \forall (x,t) \in \Omega, \tag{1.1.1}$$

$$\begin{cases} u(a,t) = u(b,t) = 0, & \forall t \in (0,T), \\ u(x,0) = u_0(x) & \text{for } \sigma(x,0) > 0, \\ u(x,T) = u_T(x) & \text{for } \sigma(x,T) < 0, \end{cases} \tag{1.1.2}$$

where $\Omega = (a,b) \times (0,T)$, σ, r, q, p, u_0, u_T are known functions, $r \geq \rho > 0$ and $-\infty < a < 0 < b < \infty$. The diffusion coefficient σ changes sign in Ω.

Problem (1.1.1), (1.1.2) arises in various applications, such as, boundary layer problems in fluid dynamics [11], [12], plasma physics and astrophysics in the study of propagation of an electron beam through the solar corona [6].

Problems like $\sigma(x,t)u_t - u_{xx} = f(x,t)$ with $\sigma(x,t)$ taking different signs were first considered by Gevrey [3], [4]. He treated in particular the case $\sigma(x,t) = x^m$ with m an odd integer. Later, in 1968, Baouendi and Grisvard [1] dealt with the case $\sigma(x,t) = x$ in detail. A similar treatment in a context where the second-order derivative is replaced by a suitable nonlinear differential operator can be found in Lions' book [7]. Franklin and Rodemich [2] considered also the case $\sigma(x,t) = x$ and treated the equation on $-\infty < x < \infty$, $0 < t < T$. In 1976, Pagani showed the existence and uniqueness of a weak solution of problem (1.1.1), (1.1.2) for a special case $\sigma = \omega(x)$, $\omega(x)\operatorname{sgn} x > 0$ for $x \neq 0$ and $u_0 = u_T = 0$. In 1984, Goldstein and Mazumdar [5] gave also an existence theorem of a weak solution of the equation (1.1.1) with $r = 1$ subject to

* This chapter is based on the papers, H. Lu and W.-Y. Wen, *Solution of a forward-backward heat equation*, Report 9439, October 1994, Department of Mathematics, University of Nijmegen, The Netherlands (submitted).

$$\text{(1.1.3)} \qquad \begin{cases} u(a,t) = u(b,t) = 0 & \forall t \in (0,T), \\ u(x,0) = u_0(x) & \forall x \in (0,b) \\ u(x,T) = u_T(x) & \forall x \in (a,0) \end{cases}$$

under certain assumptions. Unfortunately, the problem (1.1.1), (1.1.3) is overdetermined in general even if their assumptions are satisfied as mentioned in §1.5.

The aim of the present paper is to show the existence and uniqueness of a weak solution for the problem (1.1.1), (1.1.2) in a certain Hilbert space and show that the problem is well-posed in a certain sense. Comparisons with previous result, i.e., Pagani's result [10], on the existence and uniqueness of solution of (1.1.1), (1.1.2). The results in this paper have the following advantages:

- The new result is applicable to a much wider class of equation than Pagani's result. The assumption made on the coefficients of equation in this paper is weaker that Pagani's assumption even if $\sigma = \omega(x)$, $\omega(x)\,\mathrm{sgn}\,x > 0$ for $x \neq 0$ and $u_0 = u_T = 0$. The new result and Pagani's result on the existence and uniqueness are given in different sense, but the new one is much stronger than the old one. Pagani's result follows immediately from the result in the present paper.
- It is well known that a common approach for a heat equation is first to apply the Galerkin method in space to reduce the equation to a set of ordinary differential equations. Then a suitable method is applied to integrate the ordinary differential equations. However, the forward-backward heat equation (1.1.1), (1.1.2) does not fit this category because the diffusion coefficient $\sigma(x,t)$ changes sign. The new result can serve a mathematical basis for a finite element method for problem (1.1.1), (1.1.2) that discretizes space and time variables simultaneously by using continuous finite element methods (see [8]), but Pagani's result fails for the purpose.

1.2. Notation

In this section we introduce notation used in the rest of the paper. Denote the boundary $\partial\Omega$ by $\Gamma_1 \cup \cdots \cup \Gamma_6$, where Γ_i are defined as follows:

$$\begin{aligned} \Gamma_1 &= \{(x,t) :\ x \in (a,b),\ t = 0,\ \sigma(x,0) \leq 0\}, \\ \Gamma_2 &= \{(x,t) :\ x = a,\ t \in (0,T),\}, \\ \Gamma_3 &= \{(x,t) :\ x \in (a,b),\ t = T,\ \sigma(x,T) < 0\}, \\ \Gamma_4 &= \{(x,t) :\ x \in (a,b),\ t = T\ \sigma(x,T) \geq 0\}, \\ \Gamma_5 &= \{(x,t) :\ x = b,\ t \in (0,T)\}, \\ \Gamma_6 &= \{(x,t) :\ x \in (a,b), t = 0,\ \sigma(x,0) > 0\}. \end{aligned}$$

Let $L^2(\Omega)$ be the standard space of square integrable functions on Ω with inner product $(\cdot,\cdot)$ defined by

$$(u,v) = \int_\Omega uv d\Omega$$

and norm $\|u\|_0 = (u,u)^{1/2}$. We use also the classical Sobolev space $H^m(\Omega)$ provided with the norm

$$\|u\|_m = \left(\sum_{|\alpha| \le m} \int_\Omega |\partial^\alpha u|^2 d\Omega \right)^{1/2}$$

The set $C^{(1,0)}(\Omega) = \{f \in C(\Omega), f_x \in C(\Omega)\}$ is a linear space with the operations $(f_1 + f_2)(x,t) = f_1(x,t) + f_2(x,t)$ and $(\alpha f)(x,t) = \alpha f(x,t)$, where $(x,t) \in \Omega$, $f : \Omega \to \mathbb{R}$ and $\alpha \in \mathbb{R}$.

For $f \in C(\Omega)$ the support of f is the closure in Ω of the set $\{(x,t) \in \Omega : f(x,t) \neq 0\}$. $C_0(\Omega)$ is the subset of those functions in $C(\Omega)$ with compact support. Similarly, we define $C_0^{(1,0)}(\Omega) = C^{(1,0)}(\Omega) \cap C_0(\Omega)$.

If $f : A \to B$ and $C \subset A$, notation $f|_C$ denotes the restriction of f to C. We define a linear space of functions on the closure $\overline{\Omega}$ as follows:

$$C^{(1,0)}(\overline{\Omega}) = \{ f|_{\overline{\Omega}} : \ f \in C_0^{(1,0)}(\mathbb{R}^2)\}.$$

On $C^{(1,0)}(\overline{\Omega})$ we define an inner product by

$$\langle f, g \rangle = \int_\Omega (f\overline{g} + f_x \overline{g}_x) d\Omega + \int_{\partial\Omega} \kappa(x,t) f \overline{g} dS,$$

where $\kappa(x,t)$ is a nonnegative function on $\partial\Omega$. In the present paper we choose

$$\kappa(x,t) = \begin{cases} |\sigma|, & \text{if } (x,t) \in \Gamma_1 \cup \Gamma_3 \cup \Gamma_4 \cup \Gamma_6, \\ 0, & \text{if } (x,t) \in \Gamma_2 \cup \Gamma_5. \end{cases}$$

Define $H^{(1,0)}(\Omega)$ to be the completion of the linear space $C^{(1,0)}(\overline{\Omega})$ with the norm $\|\cdot\|_{(1,0)} = \langle \cdot, \cdot \rangle^{1/2}$. Then $H^{(1,0)}(\Omega)$ is a Hilbert space.

Let $U = \{u \in H^{(1,0)}(\Omega) : \ u = 0 \text{ at } \Gamma_2 \cup \Gamma_5\}$. U is a Hilbert space with the norm $\|\cdot\|_{(1,0)}$. Finally, define $V = \{v \in H^1(\Omega) : \ v = 0 \text{ at } \Gamma_2 \cup \Gamma_5\}$.

1.3. Existence and Uniqueness of a Weak Solution

In this section we show the existence and uniqueness of a weak solution for the equation (1.1.1), (1.1.2) if the coefficient q, p, the right hand function f and the boundary condition $u_0(x)$ and $u_T(x)$ are real. For the complex case the result can be derived in a similar way. To this end, we need the following lemma.

LEMMA 1.3.1. *Let A, B, C and D be nonnegative numbers. if $A^2 + B^2 \le \tau(CB + D^2)$, where τ is a positive constant, then there exists a positive constant λ depending only on τ such that $A + B \le \lambda(C + D)$.*

Proof. A straightforward computation shows that

$$\begin{aligned}(A + B + D)^2 &\le 3(A^2 + B^2 + D^2) \le 3(\tau + 1)(CB + D^2) \\ &\le 3(\tau + 1)(C(A + B) + D^2) \le 3(\tau + 1)(C + D)(A + B + D).\end{aligned}$$

Therefor, $A + B \le A + B + D \le 3(\tau + 1)(C + D)$. □

THEOREM 1.3.2. *Assume that $\sigma, r, p, q, \sigma_t, r_x, r_{xx}, q_x \in H^0(\Omega)$ are bounded and $r \ge \rho > 0$. If there exists $g \in H^{(1,0)}(\Omega)$ satisfying inequality*

$$g_x - \sigma_t - r_{xx} - q_x + 2p \ge \frac{k}{2r} g^2 \tag{1.3.1}$$

with a positive constant $k > 1$, then for each $f \in H^0(\Omega)$, there exists an unique $u \in U$ such that for all $v \in V$

$$a(u,v) = r(v), \tag{1.3.2}$$

where $a(u,v)$ and $r(v)$ are defined by

$$\begin{aligned} a(u,v) &= \int_\Omega (-u(\sigma v)_t + ru_x v_x - u((r_x+q)v)_x + puv)d\Omega + \int_{\Gamma_1\cup\Gamma_4} |\sigma|uvdS, \\ r(v) &= \int_\Omega fvd\Omega + \int_{\Gamma_6} \sigma u_0(x)vdx - \int_{\Gamma_3} \sigma u_T(x)vdx. \end{aligned}$$

Furthermore, the weak solution u satisfies the following inequality

$$\begin{aligned} \|u\|_x \le C(&\|f\|_0 \\ &+ \||\sigma(x,0)|^{1/2}u_0(x)\|_{0,\Gamma_6} + \||\sigma(x,0)|^{1/2}u_T(x)\|_{0,\Gamma_3}), \end{aligned} \tag{1.3.3}$$

where C is a positive constant,

$$\|u\|_x = \left(\|u\|_0^2 + \|u_x\|_0^2 + \int_{\Gamma_1\cup\Gamma_4} |\sigma|u^2dS\right)^{1/2}$$

$$\|u\|_{0,\Gamma_6} = \left(\int_{\Gamma_6} u^2dx\right)^{1/2}, \qquad \|u\|_{0,\Gamma_3} = \left(\int_{\Gamma_3} u^2dx\right)^{1/2}$$

Proof. Denote the outward unit vector normal to $\partial\Omega$ by $n = (n_x, n_t)$. If u is a classical solution of (1.1.1) and (1.1.2), a straightforward calculation shows that

$$\begin{aligned} a(u,v) &= \int_\Omega (\sigma u_t - ru_{xx} + qu_x + pu)vd\Omega \\ &\quad - \int_{\Gamma_1\cup\Gamma_3\cup\Gamma_4\cup\Gamma_6} \sigma n_t uvdS + \int_{\Gamma_1\cup\Gamma_4} |\sigma|uvdS \\ &= \int_\Omega fvd\Omega + \int_{\Gamma_6} \sigma u_0 vdx - \int_{\Gamma_3} \sigma u_T(x)vdx, \end{aligned}$$

which implies that u is a solution of (1.3.2).

To finish our proof we define an inner product space $\overline{V} = \{v : v \in V\}$ with the inner product $\langle\cdot,\cdot\rangle$. Our task now is to prove that there exists a $u \in U$ such that (1.3.2) hold for all $v \in \overline{V}$. Using Cauchy-Schwarz inequality shows that

$$\left|\int_{\Gamma_1\cup\Gamma_4} |\sigma|uvdS\right| \le \left(\int_{\Gamma_1\cup\Gamma_4} |\sigma|u^2dS\right)^{1/2} \left(\int_{\Gamma_1\cup\Gamma_4} |\sigma|v^2dS\right)^{1/2}$$

The trace inequality (Nečas, 1967 [9], pp 84) assets the existence of a constant D such that

$$\int_{\Gamma_1\cup\Gamma_4} |\sigma|v^2dS \le D\||\sigma|^{1/2}v\|_1^2, \quad \forall v \in \overline{V}.$$

Therefore, it is straightforward to show that there exists a positive constant M such that

$$a(u,v) \le M\|u\|_{(1,0)}\|v\|_1, \qquad \forall u \in U,\ v \in \overline{V},$$

which implies that $a(u,v)$ is continuous in the first variable in $U \times \overline{V}$ since $\|v\|_1 < \infty$ for $v \in \overline{V}$. On the other hand, $a(u,v)$ is sesquilinear and $r(v)$ is linear and continuous in $\overline{V}$. Using the Lions' Projection Theorem [7], we deduce the existence

of $u \in U$ satisfying (1.3.2) for all $v \in \overline{V}$ provided we can find a positive constant μ such that

$$a(v,v) \geq \mu \|v\|^2_{(1,0)}$$

holds for all $v \in \overline{V}$.

Let $g \in H^{(1,0)}(\Omega)$ be a function such that (1.3.1) holds. A straightforward computation shows that for all $v \in \overline{V}$

$$\begin{aligned} a(v,v) &= \int_\Omega (-v(\sigma v)_t + r v_x v_x - v((r_x+q)v)_x + pvv) d\Omega \\ &\quad - \int_\Omega g v v_x d\Omega + \int_\Omega g v v_x d\Omega + \int_{\Gamma_1 \cup \Gamma_4} |\sigma| v^2 dS \\ &= \int_\Omega \Big(\frac{1}{2}(g_x - \sigma_t - r_{xx} - q_x + p)v^2 + g v v_x + r v_x^2 \Big) d\Omega + \frac{1}{2} \int_{\partial\Omega} \kappa(x,t) v^2 dS \\ &\geq \int_\Omega \Big(\frac{k}{4r} g^2 v^2 + g v v_x + r v_x^2 \Big) d\Omega + \frac{1}{2} \int_{\partial\Omega} \kappa(x,t) v^2 dS \\ &= \frac{k}{4} \int_\Omega \Big(\frac{gv}{r} + \frac{2}{k} v_x \Big)^2 r d\Omega + \int_\Omega \Big(1 - \frac{1}{k}\Big) r v_x^2 d\Omega + \frac{1}{2} \int_{\partial\Omega} \kappa(x,t) v^2 dS \\ &\geq \Big(1 - \frac{1}{k}\Big) \rho \int_\Omega u_x^2 d\Omega + \frac{1}{2} \int_{\partial\Omega} \kappa(x,t) v^2 dS \end{aligned}$$

For any $v \in \overline{V}$, Friedrichs' first inequality shows that there exists a positive constant D_1 such that

$$\int_\Omega v_x^2 d\Omega \geq D_1 \|v\|_0^2, \quad \forall v \in \overline{V}.$$

Since $k > 1$, it is readily seen that there exists a positive constant ν such that

$$\begin{aligned} a(v,v) &\geq \nu(\|v\|_0^2 + \|v_x\|_0^2) \\ &\quad + \frac{1}{2} \int_{\partial\Omega} \kappa(x,t) v^2 ds \geq \mu \|v\|^2_{(1,0)}, \end{aligned} \tag{1.3.4}$$

where $\mu \leq \min(\nu, 1/2)$.

Now we show the uniqueness. Let $W = \{\varphi(x,t);\ \varphi \in C_0(\Omega),\ \varphi_x, \varphi_t, \varphi_{xx} \in C(\Omega)\}$ and S be a set of W which is dense in the class of weak solutions. Take a sequence $\{u^{(n)}\}$ which converges in $\|\cdot\|_{(1,0)}$ norm to a weak solution of (1.3.2) and put

$$\sigma u_t^{(n)} = r u_{xx}^{(n)} - q u_x^{(n)} - p u^{(n)} + f^{(n)}. \tag{1.3.5}$$

Then multiply both side of equation (1.3.5) by $u^{(n)}$ and integrate on Ω. It follows from the proof of existence that

$$\begin{aligned} &\nu(\|u^{(n)}\|_0^2 + \|u_x^{(n)}\|_0^2) + \frac{1}{2} \int_{\partial\Omega} k(x,t)(u^{(n)})^2 dx \\ \leq\ & \int_\Omega f^{(n)} u^{(n)} d\Omega + \||\sigma|^{1/2} u^{(n)}\|^2_{0,\Gamma_3} + \||\sigma|^{1/2} u^{(n)}\|^2_{0,\Gamma_6} \\ \leq\ & \|f^{(n)}\|_0 \|u^{(n)}\|_0 + \||\sigma|^{1/2} u^{(n)}\|^2_{0,\Gamma_3} + \||\sigma|^{1/2} u^{(n)}\|^2_{0,\Gamma_6}, \end{aligned}$$

which implies that

$$\|u^{(n)}\|_x^2 \leq 2\mu^{-1}(\|f^{(n)}\|_0 \|u^{(n)}\|_0 + \||\sigma|^{1/2} u^{(n)}\|^2_{0,\Gamma_3} + \||\sigma|^{1/2} u^{(n)}\|^2_{0,\Gamma_6}.)$$

Setting $n \to \infty$ and using Lemma 1.3.1 we obtain the inequality (1.3.3) for the weak solution u. Furthermore, (1.3.3) shows that the weak solution is unique. □

The stability criterion (1.3.3) shows that the canonical solution u depends continuously on f, u_0, u_T, and so the problem (1.1.1), (1.1.2) is well-posed in a certain sense.

COROLLARY 1.3.3. *Let $r \geq \rho > 0$. If $\sigma_t + r_{xx} + q_x - 2p \leq c < 2\pi^2\rho/(b-a)^2$, where c is a constant, then* (1.3.2) *has a solution $u \in U$.*

Proof. Without loss of generality, assume $c > 0$ and $c = h\rho$, where $0 < h < 2\pi^2/(b-a)^2$. Let d be a constant such that $h < d < 2\pi^2/(b-a)^2$. Consider $g = \sqrt{2h\rho}\tan\left(\sqrt{d/2}(x - (b+a)/2)\right)$. A simple calculation shows that

$$g_x - \sigma_t - r_{xx} - q_x + 2p \geq g_x - c = g_x - h\rho \geq \frac{k}{2\rho}g^2 \leq \frac{k}{2r}g^2,$$

where $k = \sqrt{d/h} > 1$. Theorem 1.3.2 finishes the proof. □

1.4. Comparisons with Pagani's Result

In 1976, Pagani consider a special case of $\sigma = \omega(x)$ and $\omega(x)\operatorname{sgn} x > 0$ for equation (1.1.1) subject to condition (1.1.2) with $u_0(x) = u_T(x) = 0$ under the following assumption:

Pagani's assumption

i) $\omega(x) \in C^0([a,b])$, $\omega(x)\operatorname{sgn} x > 0$, $\omega(x) = x + o(x)$ for $x \to 0$,

ii) $r \in C^2(\Omega)$, $1/\lambda \leq r \leq \lambda$ $(1 < \lambda < \infty)$,

iii) $q \in C^1(\Omega)$, $p \in C^0(\Omega)$, $|q|$, $|p| < \nu < \infty$,

iv) $p - \frac{1}{2}q_x - \frac{1}{2}r_{xx} \geq 0$ on Ω

Write equation (1.1.1) in a divergence form

$$\omega(x)u_t - (\alpha u_x + \beta u)_x - \gamma u_x - \delta u = f,$$

where $\alpha = r$, $\alpha_x + \beta - \delta = -q$, $\beta_x + \delta = -p$. Let

$$\Phi \equiv \{\phi;\ \phi \in C_0^1(\overline{\Omega}); \phi = 0 \text{ on } \Gamma_1 \cup \Gamma_2 \cup \Gamma_4 \cup \Gamma_5\}$$

Pagani's weak solution is to search a $v \in L^2([0,T]; H_0^1(a,b))$ such that for every $\phi \in \Phi$ the following equality holds

$$\int_\Omega \Big(-\omega(x)v\overline{\phi}_t + (\alpha_x v_x + \beta v)\overline{\phi}_x - (\gamma v_x + \delta v)\overline{\phi}\Big)d\Omega = \int_\Omega f\overline{\phi}d\Omega. \tag{1.4.1}$$

Under his assumption Pagani showed the existence and uniqueness of a weak solution v satisfying (1.4.1)

Let us make the following weak assumption

Assumption A

i) $\sigma = \omega(x) \in C^0([a,b])$, $r \in C^2(\Omega)$, $r \geq \rho > 0$
$q \in C^1(\Omega)$, $p \in C^0(\Omega)$, $|q|$, $|p| < \nu < \infty$,

ii) $p - \frac{1}{2}q_x - \frac{1}{2}r_{xx} \geq -\frac{\rho\pi^2}{(b-a)^2}$.

If $u_0(x) = u_T(x) = 0$ and Assumption A is satisfied, it follows from the proof of Theorem 1.3.2 and Corollary 1.3.3 that there exists an unique $v \in U$ such that for every $\phi \in V$ (1.4.1) holds. Since Assumption A is weaker than Pagani's assumption and Φ is a subspace of V, Pagani's result follows immediately. Therefore, all results on regularity of a weak solution in Pagani's sense in [10] hold for the weak solution in the present paper.

Let u be a weak solution in the sense of this paper and v be a weak solution in Pagani's sense with $\sigma = \omega(x)$ and $u_0(x) = u_T(x) = 0$. In general $u \in V$, but $v \notin \Phi$. Therefore, the weak formulation (1.3.2) can be used for finite element computation for the weak solution. The existence and uniqueness of the weak solution provides a mathematical basis for a finite element method. The finite element method based on these results is shown to be one of the most efficient numerical methods for forward-backward heat equations [8]. The weak formulation (1.4.1) and Pagani's existence and uniqueness cannot serve the same purpose even if $\sigma = \omega(x)$ with $\omega(x)\operatorname{sgn} x > 0$, $u_0(x) = u_T(x) = 0$ and Pagani's assumption is satisfied simply because $v \notin \Phi$.

1.5. A Note on Goldstein and Mazumdar's Paper

In 1984, Goldstein and Mazumdar [5] showed the existence for a weak solution of the forward-backward heat equation (1.1.1) subject to (1.1.3). Unfortunately, the problem (1.1.1), (1.1.3) is overdetermined in general even if their assumptions are satisfied.

Let us recall Goldstein and Mazumdar's result first. To prove their weak existence theorem for the forward-backward heat equation (1.1.1), (1.1.3), Goldstein and Mazumdar made the following assumptions:

Hypothesis GM. (i) p, q, $q_x \in L^\infty(\Omega)$.

(ii) σ, $\sigma_t \in L^\infty(\Omega)$; $\sigma(x,0) \geq 0$ in $R = (0,b)$; $\sigma(x,T) \leq 0$ in $L = (a,0)$.

(iii) $\sigma_t \leq c_1$ a.e. in Ω where c_1 is a constant and either

1) q is real, $c_2 = \min\{0, \text{ess inf}(p + \overline{p} - q_x)\}$ and $c_1 < 4(b-a)^{-2} + c_2$, or

2) $c_1 \leq 4(b-a)^{-2} - \sqrt{2}(b-a)^{-1}\|q\|_\infty + \min\{0, \text{ess inf}(p+\overline{p})\}$.

Let $F = L^2(\Omega)$ be a Hilbert space under the norm

$$\|u\|_F = \left(\int_0^T \int_0^b |u_x|^2 dx dt\right)^{\frac{1}{2}} = \left(\int_\Omega |u_x|^2\right)^{\frac{1}{2}}.$$

and define

$$\Phi = \{\phi \in F \cap C^1(\overline{\Omega}) : \phi(x,0) = 0 \text{ in } L,\ \phi(x,T) = 0 \text{ in } R\}.$$

Φ is an inner product space with norm

$$\|\phi\|_\Phi = \left(\|\phi\|_F^2 + \int_0^b \sigma(x,0)|\phi(x,0)|^2 dx - \int_a^0 \sigma(x,T)|\phi(x,T)|^2 dx\right)^{\frac{1}{2}}$$

Let u_0, u_T be given and suppose

$$|\sigma(\cdot,0)|^{\frac{1}{2}} u_0(\cdot) \in L^2(R), \qquad |\sigma(\cdot,T)|^{\frac{1}{2}} u_T(\cdot) \in L^2(L). \tag{1.5.1}$$

Using this notation, Goldstein and Mazumdar [5] showed their result on existence of a weak solution for the forward-backward heat equation (1.1.1), (1.1.3) as follows:

Let F^* denotes the anti-dual of F and $(\cdot,\cdot)$ denotes the conjugate duality between F^* and F. If the Hypothesis GM and (1.5.1) hold, then for each $f \in F^*$ there exists a $u \in F$ such that for all $\phi \in \Phi$,

$$E(u,\phi) = M(\phi), \tag{1.5.2}$$

where

$$E(u,\phi) = -\int_\Omega u(\sigma\overline{\phi})_t + \int_\Omega u_x\overline{\phi}_x - \int_\Omega u(q\overline{\phi})_x + \int_\Omega pu\overline{\phi},$$

$$M(\phi) = (f,\phi) + \int_0^b \sigma(x,0)u_0(x)\overline{\phi(x,0)}dx - \int_a^0 \sigma(x,T)u_T(x)\overline{\phi(x,T)}dx,$$

Using their result, they made the following remark (see Remark 3.3 in [5]):
Let ε is a positive constant such that

$$-\int_\Omega \phi(\sigma\overline{\phi})_t + \int_\Omega \phi_x\overline{\phi}_x - \int_\Omega \phi(q\overline{\phi})_x + \int_\Omega p\phi\overline{\phi} \geq \varepsilon\|\phi\|_\Phi^2$$

and $N(M) = \sup\{|M(\phi)| :\ \phi \in \Phi,\ \|\phi\|_\Phi \leq 1\}$. Then for any solution u of (1.5.2)

$$\|u\|_F \leq \varepsilon^{-1}N(M)$$

$$N(M) \leq \|f\|_{F^*} + \||\sigma(\cdot,0)|^{1/2}u_0\|_{L^2(R)} + \||\sigma(\cdot,T)|^{1/2}u_T\|_{L^2(L)}. \tag{1.5.3}$$

With using (1.5.3), They concluded that the equation (1.1.1), (1.1.3) is well-posed.

To show the first difficulty assume that u is a solution of (1.1.1) and (1.1.3) with a σ satisfying $\sigma(x,0) \leq 0$ in L and $\sigma(x,T) \geq 0$ in R. Denote

$$\begin{aligned}
S_- &= \{(x,t) :\ t = T,\ x \in L,\ \sigma(x,T) < 0\},\\
S_T &= \{(x,t) :\ t = T,\ x \in L,\ \sigma(x,T) = 0\},\\
S_+ &= \{(x,t) :\ t = 0,\ x \in R,\ \sigma(x,0) > 0\},\\
S_0 &= \{(x,t) :\ t = 0,\ x \in R,\ \sigma(x,0) = 0\},
\end{aligned}$$

$$u_0 = \begin{cases} u_{00} & \text{for } (x,t) \in S_+,\\ u_{01} & \text{for } (x,t) \in S_0, \end{cases} \qquad u_T = \begin{cases} u_{T0} & \text{for } (x,t) \in S_-,\\ u_{T1} & \text{for } (x,t) \in S_T, \end{cases}$$

Multiplying (1.1.1) by u and integrating by parts, we get

$$\begin{aligned}
&\int_\Omega \Big(\big(-\frac{1}{2}\sigma_t - \frac{1}{2}q_x + p\big)u^2 + u_x^2\Big)d\Omega\\
&= \int_\Omega fud\Omega + \frac{1}{2}\int_a^b \sigma(x,0)u(x,0)^2dx - \frac{1}{2}\int_a^b \sigma(x,T)u(x,T)^2dx\\
&\leq \|f\|\|u\| + \frac{1}{2}\int_{S_+} \sigma(x,0)u(x,0)^2dx - \frac{1}{2}\int_{S_-} \sigma(x,T)u(x,T)^2dx,
\end{aligned} \tag{1.5.4}$$

where $\|u\| = \Big(\int_\Omega u^2d\Omega\Big)^{1/2}$. It follows immediately from the Hypothesis GM that

$$-\frac{1}{2}\sigma_t - \frac{1}{2}q_x + p \geq -\frac{1}{2}c_1 + \frac{1}{2}c_2 > -\frac{2}{(b-a)^2}.$$

Using the well-known inequality

$$\int_a^b g'(x)^2dx \geq \frac{\pi^2}{(b-a)^2}\int_a^b g(x)^2dx$$

for functions satisfying $g(a) = g(b) = 0$ shows that

$$\int_\Omega u_x^2 d\Omega \geq \frac{\pi^2}{(b-a)^2} \int_\Omega u^2 d\Omega,$$

which implies that there exists a positive constant C_1 such that

$$\int_\Omega \left((-\frac{1}{2}\sigma_t - \frac{1}{2}q_x + p)u^2 + u_x^2 \right) d\Omega \geq C_1 \int_\Omega (u^2 + u_x^2) d\Omega.$$

Therefore, it follows from (1.5.4) that there is a positive constant $\widetilde{C}$ such that

$$\begin{aligned} &\int_\Omega (u^2 + u_x^2) d\Omega \\ &\leq \widetilde{C} \left(\|f\| \|u\| + \frac{1}{2} \int_{S_+} \sigma(x,0) u(x,0)^2 dx - \frac{1}{2} \int_{S_-} \sigma(x,T) u(x,T)^2 dx \right). \end{aligned}$$

Applying Lemma 1.3.1 shows that there exists a positive constant C such that

$$\begin{aligned} \|u\| + \|u_x\| \leq C \Bigg(&\|f\| + \left(\int_{S_+} \sigma(x,0) u(x,0)^2 dx \right)^{1/2} \\ &+ \left(\int_{S_-} -\sigma(x,T) u(x,T)^2 dx \right)^{1/2} \Bigg), \end{aligned}$$

which implies u is uniquely determined on $S_0 \cup S_T$ by f, u_{00} and u_{T0} if u is continuous on $\overline{\Omega}$. In particular, the canonical solution of the problem (1.1.1), (1.1.3) is the case if it has a solution. Therefore, problem (1.1.1), (1.1.3) is overdetermined if the measure of $S_0 \cup S_T$ is large than zero.

References

[1] M. S. Baouendi and P. Grisvard, *Sur une équation d'évolution changegeant de type*, J. Funct. anal., 2 (1968), pp. 352–367.

[2] J. A. Franklin and E. R. Rodemich, *Numerical analysis of an elliptic-parabolic partial differential equation*, SIAM J. Numer. Anal., 5 (1968), pp. 680–716.

[3] M. Gevrey, *Sur les équations aux dérivées partielles du type parabolique*, J. Math. pures Appl., 6 (1913), pp. 305–475.

[4] ———, *Sur les équations aux dérivées partielles du type parabolique (suite)*, J. Math. pures Appl., 6 (1914), pp. 105–148.

[5] J. A. Goldstein and T. Mazumdar, *A heat equation in which the diffusion coefficient changes sign*, J. Math. Anal. Appl., 103 (1984), pp. 533–564.

[6] T. LaRosa, *The propagation of an electron beam through the solar corona*, PhD thesis, Department of Physics and Astronomy, University of Maryland, 1986.

[7] J. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod Gauthier-Villars, Paris, 1969.

[8] H. Lu, *Galerkin and weighted Galerkin methods for the forward-backward heat equation*, tech. rep., 9410, Department of Mathematics, University of Nijmegen, The Netherlands, 1994.

[9] J. Nečas, *Les Méhtodes Directes en Théorie des Equations Elliptiques*, Masson, Paris, 1967.

[10] C. M. Pagani, *On forward-backward parabolic equations in bounded domains*, Bolletino U. M. I., (5) **13**-B (1976), pp. 336–354.

[11] K. Stewartson, *Multistructural boundary layers on flat plates and related bodies*, Adv. in Appl. Mech., 14 (1974), pp. 145–239.

[12] ———, *D'Alembert's paradox*, SIAM Rev., 23 (1981), pp. 308–343.

CHAPTER 2

Galerkin and Weighted Galerkin Methods*

Abstract. Galerkin and weighted Galerkin methods are proposed for the numerical solution of parabolic partial differential equations where the diffusion coefficient takes different signs. The approach is based on a simultaneous discretization of space and time variables by using continuous finite element methods. Under some simple assumptions, error estimates and some numerical results for both Galerkin and weighted Galerkin methods are presented. Comparisons with the previous methods show that new methods not only can be used to solve a wider class of equations but also require less regularity for the solution and need fewer computations.

Key words. Galerkin method, weighted Galerkin method, forward-backward heat equation, error estimate

AMS subject classifications. 65N30, 35K20

2.1. Introduction

In this paper we consider Galerkin and weighted Galerkin methods for the following forward-backward heat equation:

$$\sigma(x,t)u_t - u_{xx} = f(x,t), \quad \forall (x,t) \in \Omega = (-1,1) \times (0,1), \tag{2.1.1}$$

$$\begin{cases} u(\pm 1,t) = 0, & \forall t \in (0,1), \\ u(x,0) = 0 & \text{for } \sigma(x,0) > 0, \\ u(x,1) = 0 & \text{for } \sigma(x,1) < 0, \end{cases} \tag{2.1.2}$$

where the coefficient $\sigma(x,t)$ changes sign in Ω.

Problem (2.1.1), (2.1.2) arises in various applications, such as, boundary layer problems in fluid dynamics [13], [14], plasma physics and astrophysics in the study of propagation of an electron beam through the solar corona [8].

Problems like $\sigma(x,t)u_t - u_{xx} = f(x,t)$ with $\sigma(x,t)$ taking different signs were first considered by Gevrey [6], [7]. He treated in particular the case $\sigma(x,t) = x^m$ with m an odd integer. Later, in 1968, Baouendi and Grisvard [2] dealt with the case $\sigma(x,t) = x$ in detail. A similar treatment in a context where the second-order derivative is replaced by a suitable nonlinear differential operator can be found in Lions' book [9]. Franklin and Rodemich [4] considered also the case $\sigma(x,t) = x$ and treated the equation on $-\infty < x < \infty$, $0 < t < T$. Recently, Lu and Wen [10] showed the existence and uniqueness of a weak solution for the equation (2.1.1), (2.1.2) on a certain Hilbert space. The existence and uniqueness of a weak solution

* This chapter is based on the paper, H. Lu, *Galerkin and weighted Galerkin methods for a forward-backward heat equation*, Report 9410, January 1994, Department of Mathematics, University of Nijmegen, The Netherlands (submitted).

in a different sense for a special $\sigma = \sigma(x)$ satisfying $\sigma \operatorname{sgn} x > 0$ was given by Pagani in 1976 [11]. The stability criterion in [10] shows the problem is well-posed.

It is well known that a common approach for a heat equation is first to apply the Galerkin method in space to reduce the equation to a set of ordinary differential equations. Then a suitable method is applied to integrate the ordinary differential equations. However, the forward-backward heat equation (2.1.1), (2.1.2) does not fit this category because the diffusion coefficient $\sigma(x,t)$ changes sign. In 1990, Vanaja and Kellogg [15] presented some iterative methods to solve finite difference approximation to (2.1.1), (2.1.2), in which the unknowns are swept in the order suggested by the equation, if $\sigma(x,t) = \sigma(x)$ or $\sigma_t(x,t) \leq 0$. In 1991, Aziz and Liu [1] consider a finite element method for (2.1.1) subject to

$$\text{(2.1.3)} \qquad \begin{cases} u(-1,t) = u(1,t) = 0, & \forall t \in (0,1), \\ u(x,0) = 0, & \forall x \in (0,1), \\ u(x,1) = 0, & \forall x \in (-1,0). \end{cases}$$

Though the problem (2.1.1), (2.1.3) is overdetermined in general as we have seen in the previous chapter, Aziz-Liu's method can be used to solve (2.1.1), (2.1.2) with a straightforward modification. Their approaches are to transform the equation to a first-order system of symmetric-positive partial differential equations in the sense of Friedrichs [5] and to solve the system by a finite element method.

The aim of the paper is to present Galerkin and weighted Galerkin methods for (2.1.1), (2.1.2) without transforming the equation to a first-order system of partial differential equations. We consider a simultaneous discretization of space and time variables by using continuous finite element methods. If there exist two functions g, $q \in H^1(\Omega)$ such that $|q| \leq C < +\infty$ and

$$\text{(2.1.4)} \qquad \frac{1}{2} g_x - \frac{1}{2} \sigma_t - \sigma q_t - q_x^2 \geq \frac{\beta}{4} g^2,$$

which holds in particular for $\sigma_t(x,t) \leq b < \pi^2/2$, our results show that the L^2 rate of convergence is $O(h^k)$, where h is the meshsize of space and time, if the solution $u \in H^{k+1}(\Omega)$ and piecewise polynomials of degree k are used.

It is shown also that the Galerkin method works well under certain assumptions even if $\sigma_t(x,t) \geq \pi^2/2$ in some points of Ω, but there exist some $\sigma(x,t)$ for which the Galerkin method may fail to solve (2.1.1), (2.1.2). The weighted Galerkin method can be useful to solve a wide class of the forward-backward heat equations like $\sigma(x,t)u_t - u_{xx} = f(x,t)$. Some examples show that for some $\sigma(x,t)$ the weighted Galerkin method can be used to solve (2.1.1), (2.1.2) efficiently in case the Galerkin method fails. The linear systems of the discrete equations arising from the methods are positive definite. Comparisons with previous methods known for the forward-backward heat equation, for example Vanaja-Kellogg's method [15] and Aziz-Liu's method [1]. The methods presented in this paper have the following advantages:

- The new methods can be used to solve a much wider class of equations than Vanaja-Kellogg's method and Aziz-Liu method. The assumption made on the coefficients of the equation is weaker than previous ones. Aziz-Liu's assumptions are stronger than (2.1.4) if their method is used to solve (2.1.1), (2.1.2). Their assumptions imply (2.1.4). Furthermore, it is shown that if Aziz and Liu's method is applicable to (2.1.1), (2.1.2), so is the weighted Galerkin method. The difference between $\sigma_t < \pi^2/2$ and Vanaja-Kellogg's

assumption $\sigma_t \leq 0$ is essential. A example shows that doing a transformation $y = y(t)$ for a wide class of equations (2.1.1), (2.1.2) with $\sigma_t \geq \pi^2/2$ for some points in Ω, we obtain new forward-backward heat equations (2.1.1), (2.1.2) such that the corresponding σ satisfies $\sigma_t < \pi^2/2$, but there is no transformation $y = y(t)$ such that the corresponding σ satisfies $\sigma_t \leq 0$.

- The new methods require less regularity for the solution of the equations than the previous numerical methods. Vanaja-Kellogg's method requires that the solution possesses a continuous derivative of order 4 in x and order 2 in t to obtain the rate of convergence $O(k+h^2)$, where k and h are meshsize in time and in space, respectively. Aziz-Liu's method requires the solution $u_t, u_{xt}, u_{xx} \in L^2(\Omega)$ if it is used to solve (2.1.1), (2.1.2). The new methods need only the solution $u \in H^1(\Omega)$.
- Both Galerkin and weighted Galerkin methods need fewer computations than Aziz-Liu's method because they only involve half the number of unknowns.
- Unlike Aziz-Liu's methods, the new methods do not need to preprocess the boundary condition to match the boundary condition required by the corresponding first-order systems. It still remains unknown how to do such preprocessing.

At the end of the introduction we introduce some notation to be used in the paper. Denote the boundary $\partial\Omega$ by $\Gamma_1 \cup \cdots \cup \Gamma_6$, where Γ_i are defined as follows:

$$\begin{aligned}
\Gamma_1 &= \{(x,t) : \ x \in (-1,1), \ t = 0, \ \sigma(x,0) \leq 0\}, \\
\Gamma_2 &= \{(x,t) : \ x = -1, \ t \in (0,1)\}, \\
\Gamma_3 &= \{(x,t) : \ x \in (-1,1), \ t = 1, \ \sigma(x,1) < 0\}, \\
\Gamma_4 &= \{(x,t) : \ x \in (-1,1), \ t = 1, \ \sigma(x,1) \geq 0\}, \\
\Gamma_5 &= \{(x,t) : \ x = 1, \ t \in (0,1)\}, \\
\Gamma_6 &= \{(x,t) : \ x \in (-1,1), t = 0, \ \sigma(x,0) > 0\},
\end{aligned}$$

and the outward unit vector normal to $\partial\Omega$ by $n = (n_x, n_t)$.

Let $L^2(\Omega)$ be the standard space of square integrable functions on Ω with inner product $(\cdot,\cdot)$ defined by $(u,v) = \int_\Omega uv d\Omega$ and norm $\|u\|_0 = (u,u)^{\frac{1}{2}}$. We use also the classical Sobolev space $H^m(\Omega)$ provided with the norm

$$\|u\|_m = \left(\sum_{|\alpha| \leq m} \int_\Omega |\partial^\alpha u|^2 d\Omega \right)^{\frac{1}{2}}$$

and the seminorm

$$|u|_m = \left(\sum_{|\alpha| = m} \int_\Omega |\partial^\alpha u|^2 d\Omega \right)^{\frac{1}{2}}$$

Finally, define the test and trial space by

$$V = \{v \in H^1(\Omega) : \ v = 0 \text{ at } \Gamma \equiv \Gamma_2 \cup \Gamma_3 \cup \Gamma_5 \cup \Gamma_6\}.$$

and for $u \in H^1(\Omega)$ define

$$\|u\|_x = (\|u\|_0^2 + \|u_x\|_0^2)^{1/2}, \qquad \|u\|_{x,w} = \|wu\|_x,$$

where $w \in H^1(\Omega)$ is a positive function.

2.2. A Galerkin Variational Formulation

Define $a(u,v) = (\sigma u_t, v) + (u_x, v_x)$. The Galerkin variational formulation of equation (2.1.1) for a given $f \in L^2(\Omega)$ is to find a $u \in V$ such that

$$a(u,v) = (f,v), \quad \forall v \in V, \tag{2.2.1}$$

Note that for any $u \in H^2(\Omega) \cap V$ it is readily seen that

$$a(u,v) = (\sigma u_t, v) - (u_{xx}, v).$$

THEOREM 2.2.1. *If there exists a function $g \in H^1(\Omega)$ such that*

$$g_x - \sigma_t \geq \frac{\beta}{2} g^2 \tag{2.2.2}$$

with $\beta > 1$, then there exists a positive constant C such that

$$\|u\|_x^2 \leq C a(u,u), \quad \forall u \in V. \tag{2.2.3}$$

Proof. Friedrichs' first inequality shows that there exists a positive constant γ such that

$$\int_\Omega u_x^2 d\Omega \geq \gamma \|u\|_0^2, \quad \forall u \in V,$$

which implies that (2.2.3) holds if and only if there exists a positive $\widetilde{C}$ such that

$$\int_\Omega u_x^2 d\Omega \leq \widetilde{C} a(u,u). \tag{2.2.4}$$

On the other hand, a simple computation shows that

$$\begin{aligned} a(u,u) &= \int_\Omega (\sigma u_t u + u_x^2 - g u_x u + g u_x u) d\Omega \\ &= \int_\Omega \Big(\frac{1}{2}(\sigma u^2)_t - \frac{1}{2}\sigma_t u^2 + u_x^2 - \frac{1}{2}(g u^2)_x + \frac{1}{2} g_x u^2 + g u_x u \Big) d\Omega \\ &= \int_\Omega \Big(\frac{1}{2}(g_x - \sigma_t) u^2 + g u_x u + u_x^2 \Big) d\Omega + \int_\Omega \Big(\frac{1}{2}(\sigma u^2)_t - \frac{1}{2}(g u^2)_x \Big) d\Omega. \end{aligned}$$

With the use of Green's formula, the last integration turns out to be

$$\begin{aligned} &\int_\Omega \Big(\frac{1}{2}(\sigma u^2)_t - \frac{1}{2}(g u^2)_x \Big) d\Omega = \int_{\partial\Omega} \Big(\frac{1}{2}\sigma n_t u^2 - \frac{1}{2} g n_x u^2 \Big) dS \\ = &\int_{\Gamma_1 \cup \Gamma_4} \Big(\frac{1}{2}\sigma n_t u^2 - \frac{1}{2} g n_x u^2 \Big) dS = \int_{\Gamma_1 \cup \Gamma_4} \frac{1}{2}\sigma n_t u^2 dS \geq 0. \end{aligned}$$

Hence, assuming that g is a function satisfying (2 2.2), we have

$$\begin{aligned} a(u,u) &\geq \int_\Omega \Big(\frac{\beta}{4} g^2 u^2 + g u_x u + u_x^2 \Big) d\Omega \\ &= \int_\Omega (\beta(g u/2 + u_x/\beta)^2 + (1 - 1/\beta) u_x^2) d\Omega \\ &\geq \widetilde{C}^{-1} \int_\Omega u_x^2 d\Omega, \end{aligned}$$

where $1/\widetilde{C} = 1 - 1/\beta$. □

Applying Theorem 2.2.1, we have the following nice simple condition for (2.2.3).

COROLLARY 2.2.2. *Suppose $\sigma_t \le b < \pi^2/2$, where b is a positive constant. Then*

$$\|u\|_x^2 \le Ca(u,u), \quad \forall u \in V, \tag{2.2.5}$$

where C is a positive constant only depending on b.

Proof. Since $b < \pi^2/2$, there exists a positive constant c such that $b < c < \pi^2/2$. Consider the function $g(x) = \sqrt{2b}\tan\sqrt{\frac{c}{2}}x$. A simple computation shows that

$$\begin{aligned} g_x - \sigma_t &\ge g_x - b = \sqrt{bc}\tan^2\sqrt{\frac{c}{2}}x + \sqrt{b}(\sqrt{c} - \sqrt{b}) \\ &= \frac{1}{2}\sqrt{\frac{c}{b}}g^2 + \sqrt{b}(\sqrt{c} - \sqrt{b}). \end{aligned}$$

Applying Theorem 2.2.1 finishes the proof. □

The following example shows that the difference between $\sigma_t < \pi^2/2$ and Vanaja-Kellogg's assumption $\sigma_t \le 0$ is essential.

Example 1. Let $\sigma = e^t\theta(x) - \varphi(x)$, where $\theta(x)$ and $\varphi(x)$ are a continuous function for $-1 \le x \le 1$ such that σ changes sign in Ω and $\varphi \le b < \pi^2/2$. Doing a transformation $y = (1 - e^{-t})/(1 - e^{-1})$ for (2.1.1), (2.1.2), we obtain a new forward-backward heat equation

$$\alpha(x,y)v_y - v_{xx} = \widetilde{f}(x,y), \qquad \forall (x,y) \in (-1,1) \times (0,1)$$

subject to condition (2.1.2), where $v(x,y) = u(x,t)$, $\widetilde{f}(x,y) = f(x,t)$ and $\alpha(x,y) = (\theta(x) - \varphi(x))/(1 - e^{-1}) + \varphi(x)y$. Therefore, $\alpha_y(x,y) = \varphi(x) \le b < \pi^2/2$.

Denote the corresponding σ by α after transformation $y = y(t)$. Now we show that there is no transformation $y = y(t)$ such that $\sigma_t \le 0$ in general. A straightforward computation shows

$$\alpha_y = \sigma_t + (\log y')'\sigma.$$

Hence, for the zero points of σ, $\alpha_y = \sigma_t = e^t\theta(x) = \varphi(x)$, which implies that there is no any transformation $y = y(t)$ such that $\alpha_t \le 0$ if $\varphi(x) > 0$ on the zero points of σ.

Inequality (2.2.3) may not be true for some σ if the conditions of Corollary 2.2.2 are not satisfied. The following example shows that there exists at least one function $u \in V$ such that $a(u,u) \le 0$ if $\inf\sigma_t \ge \pi^2/2$.

Example 2. Assume that $\inf\sigma_t \ge \pi^2/2$. Consider $u = t(t-1)\cos\frac{\pi}{2}x$. It is readily seen that

$$\begin{aligned} a(u,u) &= \int_\Omega \Big(-\frac{1}{2}\sigma_t u^2 + u_x^2\Big)d\Omega \le \int_\Omega \Big(-\frac{\pi^2}{4}u^2 + u_x^2\Big)d\Omega \\ &= -\frac{\pi^2}{2}\int_\Omega t(t-1)\Big(\cos^2\frac{\pi}{2}x - \sin^2\frac{\pi}{2}x\Big)d\Omega = 0 \end{aligned}$$

Example 2 shows that the Galerkin approximation based on (2.2.1) may fail for solving (2.1.1), (2.1.2) if the conditions of Theorem 2.2.1 are not satisfied. Our next example, however, shows that there exists a positive constant C such that (2.2.3) hold for some $\sigma \in H^1(\Omega)$ even if $\sigma_t \ge \pi^2/2$ for some points $(x,t) \in \Omega$.

Example 3. Let $\sigma = tx\tan^2 bx$, where $0 < b < \pi/2$ and $-1 \le x \le 1$. Consider $g(x) = c\tan dx$ with $\pi/2 > d > \max(\sqrt{2}, b)$ and $d - \sqrt{d^2-2} < c < d + \sqrt{d^2-2}$.

We have

$$\begin{aligned} g_x - \sigma_t &= \frac{cd}{\cos^2 dx} - x\tan^2 bx \geq \frac{cd}{\cos^2 dx} - \tan^2 bx \\ &\geq \frac{cd}{\cos^2 dx} - \tan^2 dx \geq (cd-1)\tan^2 dx \geq \frac{\beta}{2}g^2, \end{aligned}$$

where $\beta = 2(cd-1)/c^2 > 1$. Theorem 2.2.1 shows that there exists a positive constant C such that $a(u,u) \geq C\|u\|_x^2$ for $\forall u \in V$.

2.3. A Weighted Galerkin Variational Formulation

As we have seen in the previous section, the Galerkin approximation fails for solving equations like $\sigma(x,t)u_t - u_{xx} = f(x,t)$ in some cases. In this section, we will introduce a weighted Galerkin variational formulation for (2.1.1), (2.1.2) to solve a wide class of the equations. To this end, let $w(x,t) \in H^1(\Omega)$ be a function such that $k_1 \leq w(x,t) \leq k_2$, where k_1 and k_2 are positive constants, and $W(u,v) = (\sigma w u_t, v) + (w_x u_x, v) + (w u_x, v_x)$. Our weighted variational formulation of equation (2.1.1) is to find $u \in V$ such that

$$W(u,v) = (wf, v), \quad \forall v \in V. \tag{2.3.1}$$

THEOREM 2.3.1. *If there are two functions g, $q \in H^1(\Omega)$ such that $|q| \leq a < +\infty$ and*

$$\frac{1}{2}g_x - \frac{1}{2}\sigma_t - \sigma q_t - q_x^2 \geq \frac{\beta}{4}g^2 \tag{2.3.2}$$

with $\beta > 1$, then there exists a positive function $w \in H^1(\Omega)$ such that $k_1 \leq w \leq k_2$ and

$$W(u,u) \geq C\|u\|_{x,w}^2. \tag{2.3.3}$$

where k_1, k_2 and C are positive constants.

Proof. Consider $w = \alpha^2$, where $\alpha = \exp(q)$. Then

$$W(u,u) = \int_\Omega (\sigma\alpha^2 u_t u + 2\alpha_x \alpha u_x u + \alpha^2 u_x^2) d\Omega.$$

Since

$$\begin{aligned} \alpha^2 u_t u &= (\alpha u)_t(\alpha u) - \frac{\alpha_t}{\alpha}(\alpha u)^2 = \tilde{u}_t\tilde{u} - q_t\tilde{u}^2, \\ \alpha^2 u_x^2 &= (\alpha u)_x^2 - \alpha_x^2 u^2 - 2\alpha_x\alpha u_x u = \tilde{u}_x^2 - q_x^2\tilde{u}^2 - 2\alpha_x\alpha u_x u, \end{aligned}$$

where $\tilde{u} = \alpha u$, it follows from the proof of Theorem 2.2.1 that

$$\begin{aligned} W(u,u) &= \int_\Omega (\sigma\tilde{u}_t\tilde{u} - \sigma q_t\tilde{u}^2 - q_x^2\tilde{u}^2 + \tilde{u}_x^2) d\Omega \\ &\geq \int_\Omega \Big(\Big(\frac{1}{2}g_x - \frac{1}{2}\sigma_t - \sigma q_t - q_x^2\Big)\tilde{u}^2 + g\tilde{u}_x\tilde{u} + \tilde{u}_x^2\Big) d\Omega. \end{aligned}$$

Hence, if (2.3.2) holds, we have (2.3.3). □

Theorem 2.3.1 and Corollary 2.2.2 show the following result.

COROLLARY 2.3.2. *If there exists a function $q \in H^1(\Omega)$ such that $|q| < a < +\infty$ and*

$$\frac{1}{2}\sigma_t + \sigma q_t + q_x^2 \leq b < \frac{\pi^2}{4}, \tag{2.3.4}$$

then there exist a positive function $w \in H^1(\Omega)$ *such that* $k_1 \le w \le k_2$ *and*

$$W(u,u) \ge C\|u\|^2_{x,w},$$

where k_1, k_2 *and* C *are positive constants.*

The following example shows that the conditions of Corollary 2.3.2 still hold for some σ even if (2.2.3) fails.

Example 4. Consider $\sigma = \frac{1}{2}x + bt^2$, where $\pi^2/4 \le b \le \pi^2 - 1$.

First we prove that there is no function $g \in H^1(\Omega)$ such that (2.2.2) holds. If it is not the case, we have $g_x/(\beta g^2/2 + 2bt) \ge 1$, which implies that

$$g(x,t) \ge 2\sqrt{\frac{bt}{\beta}}\tan\left(\sqrt{b\beta t}x + c(t)\right). \tag{2.3.5}$$

Since $g(x,t) \in H^1(\Omega)$ and $g(0,t) \ge \tan c(t)$, we have that $c(t)$ is bounded by

$$m\pi - \frac{\pi}{2} < c(t) < m\pi + \frac{\pi}{2},$$

where m is an integer. Let $c(t) = m\pi + \delta(t)\frac{\pi}{2}$ with $-1 < \delta(t) < 1$. (2.3.5) yields that

$$g(x,t) \ge 2\sqrt{\frac{bt}{\beta}}\tan\left(\sqrt{b\beta t}x + \delta(t)\frac{\pi}{2}\right).$$

Since $b \ge \pi^2/4$ and $\beta > 1$, there are some points $(x_0, t_0) \in (-1,1] \times [0,1]$ such that

$$\sqrt{b\beta t}x + \delta(t)\frac{\pi}{2} \to \frac{\pi}{2} - 0,$$

$$\tan\left(\sqrt{b\beta t}x + \delta(t)\frac{\pi}{2}\right) \to \frac{1}{\frac{\pi}{2} - \sqrt{b\beta t}x - \delta(t)\frac{\pi}{2}},$$

when $(x,t) \to (x_0 - 0, t_0)$ or

$$\sqrt{b\beta t}x + \delta(t)\frac{\pi}{2} \to -\frac{\pi}{2} + 0,$$

$$\tan\left(\sqrt{b\beta t}x + \delta(t)\frac{\pi}{2}\right) \to \frac{1}{-\frac{\pi}{2} - \sqrt{b\beta t}x - \delta(t)\frac{\pi}{2}},$$

when $(x,t) \to (x_0 + 0, t_0)$. For example

$$(x_0,t_0) = \begin{cases} \left(\dfrac{-\delta(\frac{1}{\beta})\frac{\pi}{2} + \frac{\pi}{2}}{\sqrt{b}}, \dfrac{1}{\beta}\right) & \text{if } \delta(\tfrac{1}{\beta}) \ge 0 \\ \left(\dfrac{-\delta(\frac{1}{\beta})\frac{\pi}{2} - \frac{\pi}{2}}{\sqrt{b}}, \dfrac{1}{\beta}\right) & \text{if } \delta(\tfrac{1}{\beta}) < 0. \end{cases}$$

This contradicts to that $g \in H^1(\Omega)$. In particular, it is easy to find a function $u \in V$ such that $a(u,u) \le 0$ if $b \ge \pi^2/2$. For example, $u = t(t-1)\cos\frac{\pi}{2}x$.

Now we prove that there is a function $q \in H^1(\Omega)$ such that (2.3.4) holds. Consider $q = -\sigma^2$. A computation shows that

$$\begin{aligned} &\frac{1}{2}\sigma_t + \sigma q_t + q_x^2 = bt - 4bt\sigma^2 + \sigma^2 \\ \le\ &\sqrt{\frac{b}{2}}(2\sigma+1)^{\frac{1}{2}}(1 - 4\sigma^2) + \sigma^2. \end{aligned}$$

Note that $\sigma > -\frac{1}{2}$. Let $r(x) = \sqrt{\frac{b}{2}}(2x+1)^{\frac{1}{2}}(1-4x^2) + x^2$ with $x > -\frac{1}{2}$. Then

$$r'(x) = -\sqrt{\frac{b}{2}}(2x+1)^{\frac{1}{2}}(10x-1) + 2x.$$

If $x \geq \frac{1}{2}$, $r'(x) \leq -\sqrt{b}(10x-1) + 2x < 0$. Hence

$$\begin{aligned}
\max_{x>-\frac{1}{2}} r(x) &= \max_{-\frac{1}{2}<x<\frac{1}{2}} r(x) \leq \max_{-\frac{1}{2}<x<\frac{1}{2}} \sqrt{\frac{b}{2}}(2x+1)^{\frac{1}{2}}(1-4x^2) + \frac{1}{4} \\
&= \sqrt{\frac{b}{2}}(2\cdot 10^{-1}+1)^{\frac{1}{2}}(1-4\cdot 10^{-2}) + \frac{1}{4} \\
&\leq \sqrt{\frac{\pi^2-1}{2}}\left(\frac{6}{10}\right)^{\frac{1}{2}}\left(\frac{24}{25}\right) + \frac{1}{4} \\
&< \frac{\pi^2}{4}
\end{aligned}$$

This example implies that the weighted method based on (2.3.1) can be useful for the solution of a wide class of the problems of the type $\sigma(x,t)u_t - u_{xx} = f(x,t)$ where $\sigma(x,t)$ changes sign in Ω in case the Galerkin method based on (2.2.1) fails for the purpose. Therefore, the weighted variational formulation (2.3.1) is an essential generalization of the variational formulation (2.2.1) for equation (2.1.1), (2.1.2).

2.4. Galerkin Approximations and Discretization Error Estimates

In this section, we will discretize our finite element schemes and derive L^2 error estimates for the Galerkin approximations. The Galerkin variational formulation can be viewed as a special case $w = 1$ of the weighted Galerkin variational formulation.

Let V^h be a finite-dimensional subspace of Hilbert space V satisfying the boundary condition $u|_\Gamma = 0$, where $\Gamma = \Gamma_2 \cup \Gamma_3 \cup \Gamma_5 \cup \Gamma_6$. The weighted Galerkin approximation of the equation (2.1.1) is to find a $u^h \in V^h$ such that

$$W(u^h, v^h) = (wf, v^h), \quad \forall v^h \in V^h. \tag{2.4.1}$$

THEOREM 2.4.1. *Assume the conditions of* Theorem 2.3.1 *hold. Then there exists a unique* $u^h \in V^h$ *satisfying* (2.4.1). *Moreover,*

$$\|u^h\|_{x,w} \leq C\|f\|_0, \tag{2.4.2}$$

where C *is a positive constant.*

Proof. Let $\{\phi_j\}$ be a basis for V^h and denote $u^h = \sum u_j\phi_j$, $\mathbf{u} = (u_1, \ldots, u_n)^T$ and $\mathbf{b} = (b_1, \ldots, b_n)^T$, where $b_i = (wf, \phi_i)$. Then $\mathbf{u}$ is the solution of the following linear system

$$A\mathbf{u} = \mathbf{b}, \tag{2.4.3}$$

where $A = (a_{ij})_{i,j=1}^n$ with $a_{ij} = W(\phi_j, \phi_i)$. It follows from (2.3.3) that A is a positive definite matrix. Hence (2.4.3) has a unique solution. On the other hand, Theorem 2.3.1 shows

$$\|u^h\|_{x,w}^2 \leq C_1 W(u^h, u^h) = C_1(wf, u^h) \leq C\|f\|_0\|u^h\|_{x,w},$$

which implies (2.4.2). □

It is shown in [10] that the inequality (2.4.2) holds for the canonical solution u of (2.1.1), (2.1.2), i.e., $\|u\|_{x,w} \leq \widetilde{C}\|f\|_0$, where $\widetilde{C}$ is a positive constant. Therefore, both the canonical solution of (2.1.1), (2.1.2) and our finite element solution depend continuously on the right hand side function of the equation (2.1.1).

The linear system (2.4.3) of the discrete equations arising from the method is positive definite. We have a large number of efficient algorithms to solve it, for example, generalizations of the conjugate gradient method [3], [12] showing monotone convergence for positive definite linear systems. One can also use sparse LU-factorization for the system.

We now derive L^2 error estimates for the Galerkin approximation (2.4.1).

THEOREM 2.4.2. *Let u and u^h be solutions of problems* (2.3.1) *and* (2.4.1), *respectively. If conditions of* Theorem 2.3.1 *hold, then there exists a positive constant C such that*

$$\|u - u^h\|_{x,w} \leq C \inf_{v^h \in V^h} \|w(u - v^h)\|_1. \tag{2.4.4}$$

Proof. For a given $v^h \in V^h$ Theorem 2.3 1 shows that

$$\begin{aligned} C_1\|u^h - v^h\|_{x,w}^2 &\leq W(u^h - v^h, u^h - v^h) \\ = \; W(u - v^h, u^h - v^h) &\leq C_2\|w(u - v^h)\|_1\|u^h - v^h\|_{x,w}. \end{aligned}$$

Setting $C_3 = C_2/C_1$ shows that $\|u^h - v^h\|_{x,w} \leq C_3\|w(u - v^h)\|_1$. Since

$$\|u - u^h\|_{x,w} \leq \|w(u - v^h)\|_1 + \|u^h - v^h\|_{x,w},$$

choosing $C = C_3 + 1$ shows our result. □

To analyze the error of our method, we make the following assumptions:

1. There is an $s \geq 0$ such that $u \in V \cap H^s(\Omega)$.
2. $\{V^h\}_{h>0}$ is a regular family of finite elements, where V^h is a subspace of V consisting of piecewise polynomials of degree k with $k \leq s - 1$.

Now we have the error estimate as follows:

THEOREM 2.4.3. *If conditions of* Theorem 2.3.1 *and assumptions 1, 2 hold, then there exists a positive constant $C > 0$ such that*

$$\|u - u^h\|_{x,w} \leq Ch^k|u|_{k+1}. \tag{2.4.5}$$

Proof. The theorem follows from Theorem 2.4.2 and the usual interpolation theoretic result. □

COROLLARY 2.4.4. *If conditions of* Theorem 2.3.1 *and assumptions 1, 2 hold, then there exists a positive constant $C > 0$ such that*

$$\|u - u^h\|_0 \leq Ch^k|u|_{k+1}. \tag{2.4.6}$$

Proof. Since $w > k_1 > 0$,

$$\|u - u^h\|_0 \leq \frac{\|w(u - u^h)\|_0}{k_1} \leq \frac{\|u - u^h\|_{x,w}}{k_1}.$$

The corollary follows from Theorem 2.4.3. □

2.5. Comparisons with Aziz-Liu's Method

In 1991, Aziz and Liu [1] presented a finite element method for the problem (2.1.1), (2.1.3) by reducing the equation to a first order system of partial differential equations under the following assumptions

$$
\begin{aligned}
&H1:\ \lambda\alpha\sigma - \frac{1}{2}(\alpha\sigma)_t + \frac{1}{2}(\gamma\sigma)_x \geq k_1,\\
&H2:\ \alpha \geq k_2,\\
&H3:\ \alpha_x + \gamma\sigma < 2\sqrt{k_1 k_2}\\
&H4:\ \sigma n_t|_{\Upsilon_1\cup\Upsilon_4} \geq 0,
\end{aligned}
$$

where $k_1 > 0$ and $k_2 > 0$ are positive constants, $\Upsilon_1 = \{(x,0) : -1 < x < 0\}$ and $\Upsilon_4 = \{(x,1) : 0 < x < 1\}$.

Though the problem (2.1.1), (2.1.3) is overdetermined in general, their method can be used to solve (2.1.1), (2.1.2) with a straightforward modification under $H1$–$H3$. In fact, by a transformation

$$
\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad u_1 = e^{-\lambda t}u, \quad u_2 = e^{-\lambda t}u_x,
$$

it follows from [1] that equation (2.1.1) may be written as the symmetric first-order system

$$
A_1\mathbf{u}_x + A_2\mathbf{u}_t + A_3\mathbf{u} = \mathbf{f},
$$

where

$$
A_1 = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} \lambda\sigma & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} e^{-\lambda t}f \\ 0 \end{pmatrix}.
$$

Because of this, to solve the forward-backward equation Aziz and Liu [1] presented a Galerkin method to find a vector-valued function $\mathbf{u} = (u_1, u_2)^T : \Omega \to \mathbb{R}^2$, which is a solution of the first-order system

$$
A_1\mathbf{u}_x + A_2\mathbf{u}_t + A_3\mathbf{u} = \mathbf{g} \ \text{ in } \Omega \tag{2.5.1}
$$

with the boundary condition $M\mathbf{u} \equiv u_1 = 0$ on $\Gamma_2\cup\Gamma_3\cup\Gamma_5\cup\Gamma_6$, where the function $\mathbf{g} = (f_1, f_2)^T \in (L^2(\Omega))^2$.

Define a 2×2 matrix-valued function T by $T\mathbf{v} = \begin{pmatrix} \alpha & 0 \\ \gamma\sigma & \alpha \end{pmatrix}\mathbf{v}$, where α and γ are known function to be specified such that T is bounded, and define a function space $\overline{V}$ by $\overline{V} = \{\mathbf{u} \in (H^1(\Omega))^2 : M\mathbf{u} = 0\}$. Let $B : \overline{V} \times \overline{V} \to \mathbb{R}$ by $B(\mathbf{u},\mathbf{v}) = \langle L\mathbf{u}, T\mathbf{v}\rangle$, where $\langle\cdot,\cdot\rangle$ denotes the $(L^2(\Omega))^2$ inner product. Their weak formulation of (2.5.1) for a given $\mathbf{g} \in (L^2(\Omega))^2$ is: to find a $\mathbf{u} \in \overline{V}$ such that

$$
B(\mathbf{u},\mathbf{v}) = \langle \mathbf{g}, T\mathbf{v}\rangle, \quad \forall \mathbf{v} \in \overline{V}. \tag{2.5.2}
$$

If $H1$–$H3$ hold, following [1] shows that Aziz-Liu's fundamental result that there exists a positive constant C such that the basic condition

$$
\|\mathbf{u}\|_0^2 \leq CB(\mathbf{u},\mathbf{u})
$$

holds for problem (2.1.1), (2.1.2).

First we show that if $H1$–$H3$ hold then there exist two functions q and g such that (2.3.2) holds. If follows from $H3$ that there is a positive constant $\beta > 1$ such

that $\beta(\gamma\alpha)^2 \le 4k_1k_2 - 2\gamma\sigma\alpha_x - \alpha_x^2$. Let $q = \frac{1}{2}\log\alpha - \lambda t$ and $g = \gamma\sigma/\alpha$. A straightforward computation shows that

$$\frac{1}{2}g_x - \frac{1}{2}\sigma_t - \sigma q_t - q_x^2 = \frac{1}{2}\Big(\frac{(\gamma\sigma)_x}{\alpha} - \frac{\gamma\sigma\alpha_x}{\alpha^2}\Big) - \frac{1}{2}\sigma_t - \frac{1}{2}\frac{\sigma\alpha_t}{\alpha} + \lambda\sigma - \frac{1}{4}\frac{\alpha_x^2}{\alpha^2}$$
$$= \frac{\lambda\alpha\sigma - \frac{1}{2}(\alpha\sigma)_t + \frac{1}{2}(\gamma\sigma)_x}{\alpha} - \frac{2\gamma\sigma\alpha_x + \alpha_x^2}{4\alpha^2} \ge \frac{k_1}{\alpha} - \frac{k_1k_2}{\alpha^2} + \frac{\beta}{4}\frac{(\gamma\alpha)^2}{\alpha^2} \ge \frac{\beta}{4}g^2.$$

Second, let $u_1 = e^{-\lambda t}u$, $u_2 = e^{-\lambda t}u_x$. Then $u_{1_x} = u_2$. If $\mathbf{v} = (v, 0)^T$, where $v \in V$, we have

$$\begin{aligned}
B(\mathbf{u},\mathbf{v}) &= \int_\Omega (\sigma\alpha u_{1_t}v - \alpha u_{2_x}v + \lambda\sigma\alpha u_1 v - \beta\sigma u_{1_x}v + \beta\sigma u_2 v)d\Omega \\
&= \int_\Omega (\alpha\sigma u_{1_t}v + \lambda\sigma\alpha u_1 v - \alpha u_{2_x}v)d\Omega \\
&= \int_\Omega (\sigma\alpha e^{-\lambda t}(-\lambda u + u_t)v + \lambda\sigma\alpha e^{-\lambda t}uv - \alpha e^{-\lambda t}u_{xx}v)d\Omega \\
&= \int_\Omega (\sigma\alpha e^{-\lambda t}u_t v - \alpha e^{-\lambda t}u_{xx}v)d\Omega \\
&= \int_\Omega (\sigma\alpha e^{-\lambda t}u_t v + \alpha_x e^{-\lambda t}u_x v + \alpha e^{-\lambda t}u_x v_x)d\Omega - \int_{\partial\Omega} \sigma e^{-\lambda t}n_x u_x v dS \\
&= \int_\Omega (\sigma\alpha e^{-\lambda t}u_t v + \alpha_x e^{-\lambda t}u_x v + \alpha e^{-\lambda t}u_x v_x)d\Omega
\end{aligned}$$

In this case, choosing $w = e^{-\lambda t}\alpha$, we have

$$B(\mathbf{u},\mathbf{v}) = W(u,v), \quad \forall u \in V \cap H^2(\Omega) \text{ and } v \in V,$$

which implies that if Aziz and Liu's method is applicable to solve (2.1.1), (2.1.2), so is the weighted Galerkin method presented here. On the other hand, the weighted Galerkin method needs less computations than Aziz-Liu's method, simply because the weighted Galerkin approximation is done on $L^2(\Omega)$ while the later one is on $(L^2(\Omega))^2$. If the weak formulation (2.5.2) is used to solve the forward-backward heat equation (2.1.1), (2.1.2), it requires that $u, u_t, u_{xt}, u_{xx} \in L^2(\Omega)$. Our variational formulations (2.2.1) and (2.3.1) need $u \in H^1(\Omega)$. Finally, both methods share the same rate of convergence if they converge for (2.1.1), (2.1.2) in theory.

2.6. Numerical Examples

In this section we implement our methods for some particular examples. We performed all of the following experiments by using triangular elements and piecewise linear functions as our basis. First, we consider the example implemented in [1].

Example 5. Consider the problem (2.1.1), (2.1.2) with $\sigma(x,t) = x$ and

$$f(x,t) = \begin{cases} 2x(x^2-1)t[(t-1)^2 - 4x^2 + t(t-1)] \\ \quad -2t^2[(t-1)^2 - 24x^2 + 4], & \forall x \ge 0,\ t \in [0,1], \\ 2x(x^2-1)(t-1)(2t^2 - t - 4x^2) \\ \quad -2(t-1)^2(t^2 - 24x^2 + 4), & \forall x < 0,\ t \in [0,1], \end{cases}$$

where f has been chosen so that

$$u(x,t) = \begin{cases} (x^2-1)t^2[(t-1)^2-4x^2], & \forall x \geq 0,\ t \in [0,1], \\ (x^2-1)(t^2-4x^2)(t-1)^2, & \forall x < 0,\ t \in [0,1]. \end{cases}$$

For this example it is straightforward to show that

$$\begin{aligned} \Gamma_1 &= \{(x,t):\ x \in (-1,0),\ t=0\}, \\ \Gamma_2 &= \{(x,t):\ x=-1,\ t \in (0,1)\}, \\ \Gamma_3 &= \{(x,t):\ x \in (-1,0),\ t=1\}, \\ \Gamma_4 &= \{(x,t):\ x \in (0,1),\ t=1\}, \\ \Gamma_5 &= \{(x,t):\ x=1,\ t \in (0,1)\}, \\ \Gamma_6 &= \{(x,t):\ x \in (0,1), t=0\}. \end{aligned}$$

In Table 1 we give some numerical results that show the performance of the Galerkin method.

TABLE 1. Galerkin method with $\sigma(x,t) = x$

h	max $\lvert e\rvert$	L^2 error	L^2 rate
$\frac{1}{2}$	0.5180	0.2456	
			1.4340
$\frac{1}{4}$	0.3237	0.0909	
			1.7301
$\frac{1}{8}$	0.1480	0.0274	
			1.9356
$\frac{1}{12}$	0.0849	0.0125	
			2.0655
$\frac{1}{16}$	0.0548	0.0069	
			2.2244
$\frac{1}{24}$	0.0285	0.0028	
			2.1696
$\frac{1}{32}$	0.0176	0.0015	

Let u be the solution of the forward-backward heat equation (2.1.1), (2.1.2) and u^h be the solution of the problem (2.4.1). The L^2 error in the Table 1 and throughout this section is defined by

$$L^2 \text{ error} = \|u - u^h\|_0 = \left(\int_\Omega |u-u^h|^2 d\Omega\right)^{\frac{1}{2}}.$$

Subdividing Ω into squares, Aziz and Liu [1] performed their method with piecewise bivariate polynomials of degree≤ 2 as basis. Table 2 shows their numerical results.

In Table 1 and Table 2, we can see the L^2 error and the L^2 rate of convergence for various meshsize h of the Galerkin method and Aziz-Liu's method, respectively. The corresponding max$|e|$ and the L^2 error in Table 1 are smaller than that in Table 2. The corresponding L^2 rate becomes approximately the same after $h = 1/8$. This example shows that the Galerkin method in the present paper with a basis of piecewise linear functions achieves at least the same accuracy as Aziz and Liu's method with a basis of piecewise bivariate polynomials of degree≤ 2 does.

TABLE 2. Aziz-Liu's method with $\sigma(x,t)=x$

h	max $\lvert e\rvert$	L^2 error	L^2 rate
$\frac{1}{2}$	4.271	1.104	
			1.99
$\frac{1}{4}$	1.208	0.276	
			2.02
$\frac{1}{8}$	0.316	0.067	
			2.01
$\frac{1}{16}$	0.078	0.016	

The second numerical example gives some numerical results when our method is applied to the problem (2.1.1), (2.1.2) with $\sigma_t(x,t) > \pi^2/2$ in some points of Ω.

Example 6. Let $\sigma(x,t) = x\exp((1-x)^2) + t\tan^2\frac{9\pi}{20}x$ and

$$
f(x,t) = \begin{cases}
t\Big\{\cos\frac{7\pi}{5}x\Big[\frac{49\pi^2}{50}t((t-1)^2-18x^2)+18t+((t-1)(2t-1)-18x^2)\\
(x\exp((1-x)^2)-t)\Big]+\cos\frac{2\pi}{5}x\Big[\frac{2\pi^2}{25}t((t-1)^2-18x^2)+((t-1)\\
(2t-1)-18x^2)(x\exp((1-x)^2)-t)+18t\Big]+\cos\frac{\pi}{2}x\Big[\frac{\pi^2}{4}t((t-1)^2\\
-18x^2)+2(((t-1)(2t-1)-18x^2)(x\exp((1-x)^2)+t)+18t)\Big]\\
-36x\Big[\frac{7\pi}{5}\sin\frac{7\pi}{5}x+\frac{2\pi}{5}\sin\frac{2\pi}{5}x+\pi\sin\frac{\pi}{2}x\Big]\Big\}, \quad \forall x\ge 0,\ t\in[0,1],\\
(t-1)\Big\{\cos\frac{7\pi}{5}x\Big[\frac{49\pi^2}{50}(t-1)((t-1)^2-18x^2)+(t(2t-1)-18x^2)\\
(x\exp((1-x)^2)-t)+18(t-1)\Big]+\cos\frac{2\pi}{5}x\Big[\frac{2\pi^2}{25}(t-1)((t-1)^2\\
-18x^2)+(t(2t-1)-18x^2))(x\exp((1-x)^2)-t)+18(t-1)\Big]\\
+\cos\frac{\pi}{2}x\Big[\frac{\pi^2}{4}(t-1)((t-1)^2-18x^2)+2((t(2t-1)-18x^2)\\
(x\exp((1-x)^2)+t)+18(t-1))\Big]\Big\}-36x(t-1)^2\Big[\frac{7\pi}{5}\sin\frac{7\pi}{5}x\\
+\frac{2\pi}{5}\sin\frac{2\pi}{5}x+\pi\sin\frac{\pi}{2}x\Big], \quad \forall x<0,\ t\in[0,1].
\end{cases}
$$

One can check the solution of the equation is given by

$$
u(x,t) = \begin{cases}
\cos\frac{\pi}{2}x\Big(\cos\frac{9\pi}{10}x+1\Big)t^2[(t-1)^2-18x^2], & \forall x\ge 0,\ t\in[0,1],\\
\cos\frac{\pi}{2}x\Big(\cos\frac{9\pi}{10}x+1\Big)(t-1)^2[t^2-18x^2], & \forall x<0,\ t\in[0,1],
\end{cases}
$$

It follows from Example 3 that there is a function $g\in H^1(\Omega)$ such that $g_x-\sigma_t\ge\frac{\beta}{2}g^2$ with $\beta>1$ for $\sigma = x\exp((1-x)^2)+t\tan^2\frac{9\pi}{20}x$. Thus, the Galerkin method is available to this example. $\Gamma_1,\dots,\Gamma_6$ are the same as those in Example 5. The numerical results in Table 3 show that the L^2 rate is approximately 2.

TABLE 3. Galerkin Method with $\sigma = x\exp((1-x)^2) + t\tan^2\frac{9\pi}{20}x$

h	max $\|e\|$	L^2 error	L^2 rate
$\frac{1}{2}$	1.2333	0.4143	
			1.5146
$\frac{1}{4}$	0.5501	0.1450	
			2.1709
$\frac{1}{8}$	0.2145	0.0322	
			2.1257
$\frac{1}{12}$	0.1199	0.0136	
			2.1628
$\frac{1}{16}$	0.0783	0.0073	
			2.1932
$\frac{1}{24}$	0.0418	0.0030	
			2.1851
$\frac{1}{32}$	0.0262	0.0016	

Finally, the following example shows that the weighted Galerkin method can be used to solve (2.1.1), (2.1.2) efficiently even if the Galerkin method fails for the purpose.

Example 7. Consider the equation (2.1.1), (2.1.2) with $\sigma(x,t) = \frac{1}{2}x + \frac{\pi^3+1}{2}t^2$ and

$$f(x,t) = \begin{cases} t\cos\frac{2\pi x}{2}\Big\{[x+(\pi^2+1)t^2]\big[(t-1)(2t-1) - 8\big(\sqrt{x^2+1}-1\big) \\ \exp\big(\sqrt{x^2+1}\big)\big] + \frac{\pi^2 t}{4}\big[(t-1)^2 - 8\big(\sqrt{x^2+1}-1\big)\exp\big(\sqrt{x^2+1}\big)\big] \\ +8t\Big[\frac{x^2}{\sqrt{x^2+1}}+1\Big]\exp\big(\sqrt{x^2+1}\big)\Big\} - 8\pi t^2 x\sin\frac{\pi x}{2}\exp\big(\sqrt{x^2+1}\big), \\ \qquad \forall x \geq 0,\ t \in [0,1], \\ (t-1)\cos\frac{2\pi x}{2}\Big\{[x+(\pi^2+1)t^2]\big[t(2t-1) - 8\big(\sqrt{x^2+1}-1\big) \\ \exp\big(\sqrt{x^2+1}\big)\big] + \frac{\pi^2(t-1)}{4}\big[t^2 - 8\big(\sqrt{x^2+1}-1\big) \\ \exp\big(\sqrt{x^2+1}\big)\big] + 8(t-1)\Big[\frac{x^2}{\sqrt{x^2+1}}+1\Big]\exp\big(\sqrt{x^2+1}\big)\Big\} \\ -8\pi t^2 x\sin\frac{\pi x}{2}\exp\big(\sqrt{x^2+1}\big), \quad \forall x < 0,\ t \in [0,1]. \end{cases}$$

The solution of the equation is given by

$$u(x,t) = \begin{cases} t^2\cos\frac{\pi x}{2}\big[(t-1)^2 - 8\big(\sqrt{x^2+1}-1\big)\exp\big(\sqrt{x^2+1}\big)\big], & \forall x \geq 0, \\ (t-1)^2\cos\frac{\pi x}{2}\big[t^2 - 8\big(\sqrt{x^2+1}-1\big)\exp\big(\sqrt{x^2+1}\big)\big], & \forall x < 0. \end{cases}$$

In this case the Galerkin may fail when it is used to solve (2.1.1), (2.1.2) as mentioned in Example 4. Now we apply the weighted Galerkin method to solve it. To make the weighted method efficient, we will seek a function $q(x,t) \in H^1(\Omega)$ such that

$$\frac{1}{2}\sigma_t + \sigma q_t + q_x^2 < \frac{\pi^2}{4}$$

and the weight $w = \exp(2q)$ is sufficient large so that the coefficient matrix of (2.4.3) is not near singular. To this end, let $q = -\frac{\sigma^2}{16}$. Similarly to Example 4, we find

$$\begin{aligned}
&\frac{1}{2}\sigma_t + \sigma q_t + q_x^2 = \frac{\pi^2+1}{2}t - \frac{\pi^2+1}{8}t\sigma^2 + \frac{\sigma^2}{256} \\
\leq\ & \max_{-1/2\leq x\leq 2}\left(\sqrt{\frac{\pi^2+1}{4}(2x+1)}\left(1-\frac{x^2}{4}\right)+\frac{x^2}{256}\right) \\
\leq\ & \sqrt{\frac{\pi^2+1}{20}(2\sqrt{21}+3)}\left(1-\frac{(\sqrt{21}-1)^2}{100}\right)+\frac{1}{64} \\
<\ & \frac{\pi^2}{4}.
\end{aligned}$$

For $\sigma = \frac{1}{2}x + \frac{\pi^2+1}{2}t^2$ it is easy to see that Γ_1, Γ_2, Γ_5, Γ_6 are the same as those in Example 5, but $\Gamma_3 = \emptyset$ and $\Gamma_4 = \{(x,1) : x \in (-1,1)\}$. In Table 4, we give some numerical results that show the performance of the weighted Galerkin method for our last example.

TABLE 4. Weighted Galerkin method with $\sigma = \frac{1}{2}x + \frac{\pi^2+1}{2}t^2$

h	max $\lvert e\rvert$	L^2 error	L^2 rate
$\frac{1}{2}$	1.6825	1.0988	
			1.2847
$\frac{1}{4}$	1.0249	0.4510	
			1.5939
$\frac{1}{8}$	0 5262	0.1494	
			1.7461
$\frac{1}{12}$	0.3233	0.0736	
			1.7804
$\frac{1}{16}$	0 2202	0.0441	
			1.8065
$\frac{1}{24}$	0.1223	0.0212	
			1.8363
$\frac{1}{32}$	0 0781	0.0125	

Acknowledgments

I am grateful to Professor O. Axelsson for valuable comments on the manuscript.

References

[1] A. K. AZIZ AND J.-L. LIU, *A Galerkin method for the forward-backward heat equation*, Math. Comp., 56 (1991), pp. 35–44.

[2] M. S. BAOUENDI AND P. GRISVARD, *Sur une équation d'évolution changeant de type*, J. Funct. anal., 2 (1968), pp. 352–367.

[3] S. C. EISENSTAT, H. C. ELMAN, AND M. H. SCHULTZ, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal, 20 (1983), pp. 345–357.

[4] J. A. FRANKLIN AND E. R. RODEMICH, *Numerical analysis of an elliptic-parabolic partial differential equation*, SIAM J. Numer. Anal., 5 (1968), pp. 680–716.

[5] K. O. FRIEDRICHS, *Symmetric positive differential equations*, Comm. Pure Appl. Math., 11 (1958), pp. 333–418.

[6] M. GEVREY, *Sur les équations aux dérivées partielles du type parabolique*, J. Math. pures Appl., 6 (1913), pp. 305–475.

[7] ———, *Sur les équations aux dérivées partielles du type parabolique (suite)*, J. Math. pures Appl., 6 (1914), pp. 105–148.

[8] T. LAROSA, *The propagation of an electron beam through the solar corona*, PhD thesis, Department of Physics and Astronomy, University of Maryland, 1986.

[9] J. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod Gauthier-Villars, Paris, 1969.

[10] H. LU AND Z.-Y. WEN, *Solution of a forward-backward heat equation*, tech. rep., 9439, Department of Mathematics, University of Nijmegen, The Netherlands, 1994

[11] C. M. PAGANI, *On forward-backward parabolic equations in bounded domains*, Bolletino U. M. I., (5) **13**-B (1976), pp. 336–354.

[12] Y. SAAD, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 62 (1981), pp. 105–126.

[13] K. STEWARTSON, *Multistructural boundary layers on flat plates and related bodies*, Adv. in Appl. Mech., 14 (1974), pp. 145–239.

[14] ———, *D'Alembert's paradox*, SIAM Rev., 23 (1981), pp. 308–343.

[15] V. VANAJA AND R. B. KELLOGG, *Iterative methods for a forward-backward heat equation*, SIAM J. Numer. Anal., 27 (1990), pp. 622–635.

CHAPTER 3

A Finite Element Method and Variable Transformations*

Abstract. The global space time finite element method for the forward-backward heat equation as introduced by Lu is generalized to include a wider class of diffusion functions σ with the use of a result on the existence and uniqueness of a weak solution to such problems proven by Lu and Wen. First, the theory is extended to forward-backward equations which contain additional convection and mass terms and secondly, the conditions on the diffusion function under which a unique solution exists are further weakened via more refined estimates and transformations of the computational domain. Conducted numerical tests corroborate the obtained results.

Key words. Galerkin method, forward-backward heat equation, error estimate, variable transformation

AMS subject classifications. 65N30, 35K20

3.1. Introduction

Let $\Omega = \{(x,t),\ \mu_1(t) < x < \mu_2(t),\ 0 < t < T\}$, where $\mu_1(t)$ and $\mu_2(t)$ are continuous and piecewise smooth functions on $[0,T]$. In this paper we consider a Galerkin method for the following parabolic boundary value problem

$$\sigma u_t = u_{xx} - q u_x - p u + f, \qquad \forall (x,t) \in \Omega, \tag{3.1.1}$$

$$\begin{cases} u(\mu_1(t),t) = u(\mu_2(t),t) = 0 & \forall t \in (0,T), \\ u(x,0) = u_0(x) & \text{for } \sigma(x,0) > 0, \\ u(x,t) = u_T(x) & \text{for } \sigma(x,T) < 0, \end{cases} \tag{3.1.2}$$

where σ, q, p, f, u_0, u_T are given functions. The diffusion coefficient σ changes sign in Ω. Problem (3.1.1), (3.1.2) arises in various applications, such as, boundary layer problems in fluid dynamics [18], [19], plasma physics and astrophysics in the study of propagation of an electron beam through the solar corona [9].

Problems like $\sigma(x,t)u_t - u_{xx} = f(x,t)$ with $\sigma(x,t)$ taking different signs were first considered by Gevrey [7], [8]. He treated in particular the case $\sigma(x,t) = x^m$ with m an odd integer. Later, in 1968, Baouendi and Grisvard [4] dealt with the case $\sigma(x,t) = x$ in detail. A similar treatment in a context where the second-order derivative is replaced by a suitable nonlinear differential operator can be found in Lions' book [10]. Franklin and Rodemich [6] considered also the case $\sigma(x,t) = x$ and treated the equation on $-\infty < x < \infty$, $0 < t < T$. Recently, Lu and Wen [12]

* This chapter is based on the paper: H. Lu and J. Maubach, *A finite element method and variable transformations for a forward-backward heat equation*, (manuscript).

showed the existence and uniqueness of a weak solution for the problem (3.1.1), (3.1.2) in a certain Hilbert space and that the problem (3.1.1), (3.1.2) is well posed.

In the past few years, some authors have already paid attention to the numerical solution methods for the model problem of $\mu_1(t) = -1$, $\mu_2(t) = 1$ and $p = q = u_0 = u_T = 0$. In 1990, Vanaja and Kellogg [21] presented some iterative methods to solve finite difference approximation to (3.1.1), (3.1.2) if $\sigma(x,t) = \sigma(x)$ or $\sigma_t(x,t) \leq 0$. Aziz and Liu's numerical methods [3] can also be used to solve (3.1.1) and (3.1.2) with some straightforward modifications as mentioned in [11]. Their approaches are to transform the problem to the first-order systems of partial differential equations and to solve the systems in $(L^2(\Omega)^2)$ by a Galerkin method. Recently, Lu [11] purposed a new Galerkin method for (3.1.1), (3.1.2) based on a simultaneous discretization of space and time variables by using continuous finite element methods without transforming the problem to a first-order system. If there exists a function $g \in H^0(\Omega)$ such that $g_x \in H^0(\Omega)$ and the Riccati inequality $g_x - \sigma_t \geq kg^2/2$ holds with a positive constant $k > 1$, the method is efficient for solving the equation (3.1.1), (3.1.2). In particular, if $\sigma_t < \pi^2/2$, there exists a function g satisfying the condition.

In this paper, we generalize the Galerkin methods in [11] to the general case. The generalization is based on the result of existence and uniqueness of a weak solution for (3.1.1), (3.1.2) given by Lu and Wen [12] and possesses all advantages of Lu's methods. We present an error analysis that shows an improvement of the error estimate in [11]. For the case $\mu_1(t) = -1$, $\mu_2(t) = 1$ and $q = p = u_0 = u_T = 0$, Lu in [11] suggested a weighted Galerkin method for a wide class of the problems (3.1.1), (3.1.2). It remains a open problem how to construct a weight function in general. To solve a wide class of the equations (3.1.1), (3.1.2), we do variable transformations $x = x$ and $y = y(t)$ for the equation such that the new equation can be solved by the standard Galerkin method. We derive some conditions under which we can do the transformation to solve a wide class of equations (3.1.1), (3.1.2) and how to construct the transformations. In particular, the conditions are automatically satisfied if σ is separable, i.e., $\sigma = \kappa(x)\varphi(t)$. For another important case where $\sigma(x,t)$ is a function of $x+ct+d$, i.e., $\sigma(x,t) = \sigma(x+ct+d)$ we first apply variable transformations $y = x + ct + d$ and $t = t$. Then using a simple function transformation we obtain a forward-backward heat equation which satisfies our more general assumptions.

3.2. Galerkin Approximation

In this section we consider a finite element approximation of the equation (3.1.1), (3.1.2). First, we introduce some notation to be used throughout the paper. Denote the boundary $\partial\Omega$ by $\Gamma_1 \cup \cdots \cup \Gamma_6$, where Γ_i are defined as follows:

$$
\begin{aligned}
\Gamma_1 &= \{(x,0) : x \in (\mu_1(0), \mu_2(0))\ \sigma(x,0) \leq 0\},\\
\Gamma_2 &= \{(\mu_1(t),t) : t \in (0,T),\},\\
\Gamma_3 &= \{(x,T) : x \in (\mu_1(T), \mu_2(T)),\ \sigma(x,T) < 0\},\\
\Gamma_4 &= \{(x,T) : x \in (\mu_1(T), \mu_2(T)),\ \sigma(x,T) \geq 0\},\\
\Gamma_5 &= \{(\mu_2(t),t) : t \in (0,T)\},\\
\Gamma_6 &= \{(x,0) : x \in (\mu_1(0), \mu_2(0)),\ \sigma(x,0) > 0\},
\end{aligned}
$$

and the outward unit vector normal to $\partial\Omega$ by $n = (n_x, n_t)^T$.

Let $L^2(\Omega)$ be the standard space of square integrable functions on Ω with inner product $(\cdot,\cdot)$ defined by

$$(u,v)_\Omega = \int_\Omega uv d\Omega$$

and norm $\|u\|_{0,\Omega} = (u,u)^{\frac{1}{2}}$. We use also the classical Sobolev space $H^m(\Omega)$ provided with the norm

$$\|u\|_{m,\Omega} = \left(\sum_{|\alpha|\le m} \int_\Omega |\partial^\alpha u|^2 d\Omega \right)^{1/2}$$

and the seminorm

$$|u|_{m,\Omega} = \left(\sum_{|\alpha|=m} \int_\Omega |\partial^\alpha u|^2 d\Omega \right)^{1/2}$$

The set $C^{(1,0)}(\Omega) = \{f \in C(\Omega), f_x \in C(\Omega)\}$ is a linear space with the operations $(f_1+f_2)(x,t) = f_1(x,t) + f_2(x,t)$ and $(\alpha f)(x,t) = \alpha f(x,t)$, where $(x,t) \in \Omega$, $f : \Omega \to \mathbb{R}$ and $\alpha \in \mathbb{R}$.

For $f \in C(\Omega)$ the support of f is the closure in Ω of the set $\{(x,t) \in \Omega : f(x,t) \neq 0\}$. $C_0(\Omega)$ is the subset of those functions in $C(\Omega)$ with compact support. Similarly, we define $C_0^{(1,0)}(\Omega) = C^{(1,0)}(\Omega) \cap C_0(\Omega)$.

If $f : A \to B$ and $C \subset A$, notation $f|_C$ denotes the restriction of f to C. We define a linear space of functions on the closure $\overline{\Omega}$ as follows:

$$C^{(1,0)}(\overline{\Omega}) = \{ f|_{\overline{\Omega}} : \ f \in C_0^{(1,0)}(\mathbb{R}^2)\}.$$

On $C^{(1,0)}(\overline{\Omega})$ we define an inner product by

$$\langle f,g\rangle = \int_\Omega (fg + f_x g_x) d\Omega + \int_{\partial\Omega} \kappa(x,t) fg dS,$$

where $\kappa(x,t)$ is a nonnegative function on $\partial\Omega$. In the present paper we choose

$$\kappa(x,t) = \begin{cases} |\sigma|, & \text{if } (x,t) \in \Gamma_1 \cup \Gamma_3 \cup \Gamma_4 \cup \Gamma_6, \\ 0, & \text{if } (x,t) \in \Gamma_2 \cup \Gamma_5. \end{cases}$$

Define $H^{(1,0)}(\Omega)$ to be the completion of the linear space $C^{(1,0)}(\overline{\Omega})$ with the norm

$$\|\cdot\|_{(1,0)} = \langle \cdot,\cdot\rangle^{1/2}.$$

Then $H^{(1,0)}(\Omega)$ is a Hilbert space.

Let $U = \{u \in H^{(1,0)}(\Omega) : \ u = 0 \text{ at } \Gamma_2 \cup \Gamma_5\}$. U is a Hilbert space with the norm $\|\cdot\|_{(1,0)}$. Finally, define $V = \{v \in H^1(\Omega) : \ v = 0 \text{ at } \Gamma_2 \cup \Gamma_5\}$.

Using this notation, the variational formulation of (3.1.1), (3.1.2) for given $f,\ q,\ p \in L^2(\Omega)$, $u_0 \in L^2(\Gamma_6)$, $u_T \in L^2(\Gamma_3)$ is to find a $u \in V$ such that

$$a(u,v) = r(v), \qquad \forall v \in V, \tag{3.2.1}$$

where $a(u,v)$ and $r(v)$ are defined by

$$\begin{aligned} a(u,v) &= \int_\Omega (-u(\sigma v)_t + u_x v_x - u(qv)_x + puv) d\Omega + \int_{\Gamma_1\cup\Gamma_4} |\sigma| uv dS \\ r(v) &= \int_\Omega fv d\Omega + \int_{\Gamma_6} \sigma u_0(x) v dx - \int_{\Gamma_3} \sigma u_T(x) v dx. \end{aligned}$$

Following the proof of the existence and uniqueness of a weak solution for (3.1.1), (3.1.2) in [12] for $\mu_1(t) = a$ and $\mu_2(t) = b$ we can prove the following result.

THEOREM 3.2.1. *If there exists a function $g \in H^0(\Omega)$, $g_x \in H^0(\Omega)$ satisfying the Riccati inequality*

$$g_x - \sigma_t - q_x + 2p \geq \frac{\beta}{2} g^2 \tag{3.2.2}$$

with $\beta > 1$, then there exists an unique solution $u \in U$ satisfying the weak formulation (3.2.1) *and*

$$\|v\|^2_{(1,0)} \leq Ca(v,v), \qquad \forall v \in V, \tag{3.2.3}$$

where C is a positive constant,

Proof. The proof is essentially the same as that of Theorem 3.1 in [12]. □

COROLLARY 3.2.2. *If $\sigma_t + q_x - 2p \leq c(t) < 2\pi^2/(\mu_2(t) - \mu_1(t))^2$, where $c(t)$ is a continuous function on $[0,T]$, then the condition* (3.2.3) *holds.*

Proof. Without loss of generality assume that $c(t) > 0$ on $[0,T]$. Since $c(t)$, $\mu_1(t)$, $\mu_2(t)$ are continuous functions on $[0,T]$, there exists a continuous function $d(t)$ on $[0,T]$ such that $c(t) < d(t) < 2\pi^2/(\mu_2(t) - \mu_1(t))^2$. Consider $g = \sqrt{2c(t)} \tan(\sqrt{d(t)/2}(x - (\mu_2(t) + \mu_1(t))/2))$. A simple computation shows that

$$g_x - \sigma_t - q_x + 2p \geq g_x - c(t) > \frac{\sqrt{d(t)/c(t)}}{2} g^2 \geq \frac{\beta}{2} g^2,$$

where $\beta = \min_{0 \leq t \leq T} \sqrt{d(t)/c(t)} > 1$. □

For the case $2p - q_x \geq 0$, we also have the following result:

COROLLARY 3.2.3. *Assume $2p - q_x \geq 0$, σ, σ_t, $\sigma_{xt} \in C(\Omega)$ are bounded. If there exists a continuous function $\tau(t)$ such that $\mu_1(t) \leq \tau(t) \leq \mu_2(t)$ for $0 \leq t \leq T$ and*

1. *$(x - \tau(t))\sigma_t \geq 0$, $\sigma_{xt} \geq 0$ for $\sigma_t > 0$ and $\int_{\tau(t)}^{\mu_2(t)} \sqrt{\sigma_t} dx < \pi/\sqrt{2}$, or*
2. *$(x - \tau(t))\sigma_t \leq 0$, $\sigma_{xt} \leq 0$ for $\sigma_t < 0$ and $\int_{\mu_1(t)}^{\tau(t)} \sqrt{\sigma_t} dx < \pi/\sqrt{2}$,*

then the condition (3.2.3) *holds.*

Proof. We prove for the condition 1 only. The proof for 2 follows in a similar way. Define a function γ by

$$\gamma = \begin{cases} \sigma_t, & \text{if } x > \tau(t), \\ 0, & \text{if } x \leq \tau(t) \end{cases}$$

and consider the function $g = \sqrt{2\gamma} \tan \frac{\beta}{\sqrt{2}} \int_{\tau(t)}^{x} \sqrt{\gamma} dy$, where $\beta > 1$ is a positive constant such that $\beta \int_{\tau(t)}^{\mu_2(t)} \sqrt{\gamma} dx < \pi/\sqrt{2}$. A straightforward calculation shows that

$$g_x = \begin{cases} \dfrac{\beta \sigma_t}{\cos^2 \frac{\beta}{\sqrt{2}} \int_{\tau(t)}^{x} \sqrt{\sigma_t} dy} + \dfrac{\sigma_{xt}}{\sqrt{2}} \tan \dfrac{\beta}{\sqrt{2}} \displaystyle\int_{\tau(t)}^{x} \sqrt{\sigma_t} dy, & \text{if } x > \tau(t), \\ 0, & \text{if } x \leq \tau(t) \end{cases}$$

and g_x is continuous on Ω under the assumptions made in the corollary. Hence, we have

$$g_x - \sigma_t - q_x + 2p \geq g_x - \gamma. \tag{3.2.4}$$

If $x \leq \tau(t)$ then $g_x - \gamma = 0 \geq \frac{\beta}{2} g^2$. Otherwise, A computation shows that

$$\begin{aligned} g_x - \gamma &= \frac{\beta\sigma_t}{\cos^2 \frac{\beta}{\sqrt{2}} \int_{\tau(t)}^{x} \sqrt{\sigma_t} dy} + \frac{\sigma_{xt}}{\sqrt{2}} \tan \frac{\beta}{\sqrt{2}} \int_{\tau(t)}^{x} \sqrt{\sigma_t} dy - \sigma_t \\ &\geq \frac{\beta\sigma_t}{\cos^2 \frac{\beta}{\sqrt{2}} \int_{\tau(t)}^{x} \sqrt{\sigma_t} dy} - \beta\sigma_t + (\beta - 1)\sigma_t \\ &\geq \beta\sigma_t \tan^2 \frac{\beta}{\sqrt{2}} \int_{\tau(t)}^{x} \sqrt{\sigma_t} dy \\ &= \frac{\beta}{2} g^2, \end{aligned}$$

which leads to the desired result. □

Example 1. Consider $\sigma = Cx^n t^m - D$ and $\mu_1(t) = -b$, $\mu_2(t) = b$, $b > 0$ and $p = q = 0$, where n is an odd integer, m is a positive integer, C and D are constants such that σ changes sign in $(-b, b) \times (0, T)$. Then

$$\sigma_t = C\, m x^n t^{m-1} \leq m|C| b^n T^{m-1}.$$

Corollary 3.2.2 requires $|C| \leq \frac{\pi^2}{2mb^{n+2}T^{m-1}}$ to guarantee (3.2.3). In this case, choosing $\tau(t) = 0$ we find that conditions of Corollary 3.2.3 are satisfied. Furthermore, Corollary 3.2.3 shows a better bound $|C| \leq \left(\frac{n+2}{2}\right)^2 \frac{\pi^2}{2mb^{n+2}T^{m-1}}$.

This example shows that Corollary 3.2.3 indeed extends the result formulated in Corollary 3.2.2 in some way.

COROLLARY 3.2.4. *Assume* $2p - q_x \geq 0$, σ, $\sigma_t \geq 0$, $\sigma_{xt} \in C(\Omega)$ *are bounded. If there exists a continuous function* $\tau(t)$ *such that* $\mu_1(t) \leq \tau(t) \leq \mu_2(t)$ *for* $0 \leq t \leq T$ *and* $(x - \tau(t))\sigma_{xt} \geq 0$ *for* $(x, t) \in \Omega$ *and* $\max\left(\int_{\tau(t)}^{\mu_2(t)} \sqrt{\sigma_t} dx, \int_{\mu_1(t)}^{\tau(t)} \sqrt{\sigma_t} dx\right) < \pi/\sqrt{2}$, *then the condition* (3.2.3) *holds.*

Proof. Let $g = \sqrt{2\sigma_t} \tan \frac{\beta}{\sqrt{2}} \int_{\tau(t)}^{x} \sqrt{\sigma_t} dy$. The proof is essential the same as that of Corollary 3.2.3. □

Example 2. Consider Example 1 if n is an even integer. It is easy to show that there is no function $\tau(t)$ such that the conditions of Corollary 3.2.3 are satisfied. However, the conditions of Corollary 3.2.4 are satisfied for $\tau(t) = 0$. Corollary 3.2.4 shows a bound $C \leq \left(\frac{n+2}{2}\right)^2 \frac{\pi^2}{2mb^{n+2}T^{m-1}}$ for (3.2.3) which is larger that $\frac{\pi^2}{2mb^{n+2}T^{m-1}}$ a bound of C required by Corollary 3.2.2 for (3.2.3).

Let V^h be a finite dimensional subspace of Hilbert space V. The approximation of (3.2.1) is to find a discrete solution $u^h \in V^h$ such that

$$a(u^h, v^h) = r(v^h), \qquad \forall v^h \in V^h. \tag{3.2.5}$$

THEOREM 3.2.5. *Assume the conditions of* Theorem 3.2.1 *hold, then there exists a unique* $u^h \in V^h$ *satisfying* (3.2.5). *Moreover*

$$\|u^h\|_{(1,0)} \leq C(\|f\|_{0,\Omega} + \||\sigma|^{1/2} u_0\|_{0,\Gamma_6} + \||\sigma|^{1/2} u_T\|_{0,\Gamma_3}), \tag{3.2.6}$$

where C *is a positive constant.*

Proof. Let $\{\phi_i\}$ be a basis for V^h and denote $u^h = \sum u_i\phi_i$, $\mathbf{u} = (u_1, \ldots, u_n)^T$ and $\mathbf{b} = (b_1, \ldots, b_n)^T$, where $b_i = r(\phi_i)$. Then $\mathbf{u}$ is the solution of the following linear system

$$A\mathbf{u} = \mathbf{b}, \tag{3.2.7}$$

where $A = (a_{ij})_{i,j=1}^n$ with $a_{ij} = a(\phi_j, \phi_i)$. It follows from (3.2.3) that A is a positive definite matrix. Hence, (3.2.7) has a unique solution.

Applying Theorem 3.2.1 and Cauchy-Schwarz inequality shows that

$$\begin{aligned}
\|u^h\|^2_{(1,0)} &\leq C_1 a(u^h, u^h) = C_1 r(u^h) \\
&= C_1 \left(\int_\Omega f u^h d\Omega + \int_{\Gamma_6} \sigma u_0(x) u^h dx - \int_{\Gamma_3} \sigma u_T(x) u^h dx. \right) \\
&\leq C_1 (\|f\|_{0,\Omega}\|u^h\|_{0,\Omega} + \||\sigma|^{1/2} u_0\|_{0,\Gamma_6} \||\sigma|^{1/2} u^h\|_{0,\Gamma_6} \\
&\quad + \||\sigma|^{1/2} u_T\|_{0,\Gamma_3} \||\sigma|^{1/2} u^h\|_{0,\Gamma_3}) \\
&\leq C_1 \|u^h\|_{(1,0)} (\|f\|_{0,\Omega} + \||\sigma|^{1/2} u_0\|_{0,\Gamma_6} + \||\sigma|^{1/2} u_T\|_{0,\Gamma_3}),
\end{aligned}$$

which implies the inequality (3.2.6). □

3.3. Error Analysis

In this section, we derive error estimates for the Galerkin approximation (3.2.5).

THEOREM 3.3.1. *Let u and u^h be solutions of problem* (3.2.1) *and* (3.2.5), *respectively. If the conditions of* Theorem 3.2.1 *hold. Then there exists a positive constant C such that*

$$\|u - u^h\|_{(1,0)} \leq C \inf_{v^h \in V^h} \|u - v^h\|_{1,\Omega}. \tag{3.3.1}$$

Proof. For a given $v^h \in V^h$ Theorem 3.2.1 shows that

$$\begin{aligned}
&\|u^h - v^h\|^2_{(1,0)} \leq C_1 a(u^h - v^h, u^h - v^h) \\
\leq\ & C_1 a(u - v^h, u^h - v^h) \\
=\ & C_1 \left(\int_\Omega \left(-(u - v^h)(\sigma(u^h - v^h))_t + (u - v^h)_x (u^h - v^h)_x \right.\right. \\
& -(u - v^h)(q(u^h - v^h))_x + p(u - v^h)(u^h - v^h)) \, d\Omega \\
& \left.\left. + \int_{\Gamma_1 \cup \Gamma_4} \sigma(u - v^h)(u^h - v^h) dS \right) \right. \\
=\ & C_1 \left(\int_\Omega (\sigma(u - v^h)_t (u^h - v_h) + (u - v^h)_x (u^h - v^h)_x \right. \\
& q(u - v^h)_x (u^h - v^h) + p(u - v^h)(u^h - v^h)) \, d\Omega \\
& \left. + \int_{\Gamma_6} \sigma(u - v^h)(u^h - v^h) \right) dS - \int_{\Gamma_3} \sigma(u - v^h)(u^h - v^h) dS
\end{aligned}$$

Using Cauchy-Schwarz inequality shows that

$$\begin{aligned}
& \int_\Omega (\sigma(u - v^h)_t (u^h - v_h) + (u - v^h)_x (u^h - v^h)_x \\
& q(u - v^h)_x (u^h - v^h) + p(u - v^h)(u^h - v^h)) \, d\Omega \\
\leq\ & \eta \|u - v^h\|_{1,\Omega} (\|u^h - v^h\|^2_{0,\Omega} + \|(u^h - v^h)_x\|^2_{0,\Omega})^{1/2},
\end{aligned}$$

where η is a positive constant. On the other hand, using Cauchy-Schwarz inequality and the trace inequality (Nečas, 1967 [15], pp 84) shows that

$$\begin{aligned}
&\int_{\Gamma_6} \sigma(u - v^h)(u^h - v^h) dS \\
\leq\ & \left(\int_{\Gamma_6} \sigma(u - v^h)^2 dS \right)^{1/2} \left(\int_{\Gamma_6} \sigma(u^h - v^h)^2 dS \right)^{1/2} \\
\leq\ & \beta_1 \|u - v^h\|_{1,\Omega} \left(\int_{\Gamma_6} \sigma(u^h - v^h)^2 dS \right)^{1/2},
\end{aligned}$$

where β_1 is a positive constant Similarly, there is a positive constant β_2 such that

$$-\int_{\Gamma_3} \sigma(u - v^h)(u^h - v^h) dS \leq \beta_2 \|u - v^h\|_{1,\Omega} \left(-\int_{\Gamma_3} \sigma(u^h - v^h)^2 dS \right)^{1/2}.$$

Therefore

$$\begin{aligned}
\frac{1}{C_1} \|u - v^h\|^2_{(0\,1)} \leq\ & \eta \|u - v^h\|_{1,\Omega} \|u^h - v^h\|_{x,\Omega} \\
& + \beta_1 \|u - v^h\|_{1\,\Omega} \left(\int_{\Gamma_6} \sigma(u^h - v^h)^2 dS \right)^{1/2} \\
& + \beta_2 \|u - v^h\|_{1,\Omega} \left(-\int_{\Gamma_3} \sigma(u^h - v^h)^2 dS \right)^{1/2} \\
\leq\ & C_2 \|u - v^h\|_{1,\Omega} \|u^h - v^h\|_{(1,0)}
\end{aligned}$$

Choosing $C_3 = C_1 C_2$ shows that

$$\|u^h - v^h\|_{(1,0)} \leq C_3 \|u - v\|_{1,\Omega}$$

Again using the trace inequality (Nečas, 1967 [15], pp 84) shows that there exists a positive constant β_3 such that

$$\|v\|_{(1,0)} \leq \beta_3 \|v\|_{1,\Omega}, \qquad \forall v \in V$$

Hence,

$$\|u - v^h\|_{(1,0)} \leq \|u - v^h\|_{(1,0)} + \|u^h - v^h\|_{(1,0)} \leq \beta_3 \|u - v^h\|_{1,\Omega} + \|u^h - v^h\|_{(1,0)}.$$

Choosing $C = \beta_3 + C_3$ finishes our proof □

To analyze the error of our method, we make the following assumptions:

1 There is an $s \geq 0$ such that $u \in V \cap H^s(\Omega)$
2 $\{V^h\}_{h>0}$ is a regular family of finite elements, where V^h is a subspace of V consisting of piecewise polynomials of degree k with $k \leq s - 1$.

Now we have the error estimate as follows

THEOREM 3 3 2 *If the conditions of* Theorem 3 2 1 *and assumption 1, 2 hold, then there exists a positive constant C such that*

$$\|u - u^h\|_{(1,0)} \leq C h^k |u|_{k+1,\Omega} \tag{3 3 2}$$

Proof The theorem follows from Theorem 3 3 1 and from standard interpolation theoretical results □

3.4. Variable Transformations

In this section, we consider variable transformations for the equation. Our purpose is to search transformations such that we can solve a wide class of equation (3.1.1), (3.1.2) by using the method in §3.2 after the transformations. Without loss of generality, we assume that there is no point $t_0 \in (0,T)$ such that $\sigma(x,t_0)=0$ for all $x \in (\mu_1(t),\mu_2(t))$, because we can divide the problem (3.1.1), (3.1.2) into two subproblems

$$\sigma u_t = u_{xx} - qu_x - pu + f, \qquad \forall (x,t) \in \Omega_1,$$

$$\begin{cases} u(\mu_1(t),t) = u(\mu_2(t),t) = 0 & \forall t \in (0,t_0), \\ u(x,0) = u_0(x) & \text{for } \sigma(x,0) > 0, \end{cases}$$

where $\Omega_1 = \{(x,t) : (x,t) \in \Omega \text{ and } 0 < t < t_0\}$, and

$$\sigma u_t = u_{xx} - qu_x - pu + f, \qquad \forall (x,t) \in \Omega_2,$$

$$\begin{cases} u(\mu_1(t),t) = u(\mu_2(t),t) = 0 & \forall t \in (t_0,T), \\ u(x,T) = u_T(x) & \text{for } \sigma(x,T) < 0, \end{cases}$$

where $\Omega_2 = \{(x,t) : (x,t) \in \Omega \text{ and } t_0 < t < T\}$. According to Lu and Wen's result on existence and uniqueness [12] both subproblems have unique weak solutions. Therefore, we solve the subproblems separately if there is such a point. Let $y = y(t)$ be a function such that $y(0)=0$, $y(T) < \infty$ and $y'(t) > 0$, for $t \in (0,T)$. After the transformations $x = x$ and $y = y(t)$, the problem (3.1.1), (3.1.2) becomes

$$\alpha(x,y)v_t = v_{xx} - \tilde{q}v_x - \tilde{p}v + \tilde{f}, \quad \forall (x,t) \in \tilde{\Omega}, \tag{3.4.1}$$

$$\begin{cases} v(\mu_1(t(y)),y) = v(\mu_2(t(y)),y) = 0, & \forall y \in (0,y(T)), \\ v(x,0) = u_0(x) & \text{for } \alpha(x,0) > 0, \\ v(x,y(T)) = u_T(x) & \text{for } \alpha(x,y(T)) < 0, \end{cases} \tag{3.4.2}$$

where $\tilde{\Omega} = \{(x,y) : \mu_1(t(y)) < x < \mu_2(t(y)),\ 0 < y < y(T)\}$, $t(y)$ is the inverse of $y = y(t)$, $v(x,y) = u(x,t)$, $\alpha(x,y) = y'(t)\sigma(x,t)$, $\tilde{q}(x,y) = q(x,t)$, $\tilde{p}(x,y) = p(x,t)$ and $\tilde{f}(x,y) = f(x,t)$. A straightforward computation shows that

$$\alpha_y = \sigma_t + \sigma(\log y')'. \tag{3.4.3}$$

Our task now is to find a transformation $y = y(t)$ such that

$$\sigma_t + \sigma(\log y')' + \tilde{q}_x - 2\tilde{p} < c_1(t), \tag{3.4.4}$$

where $c_1(t) < 2\pi^2/(\mu_2(t) - \mu_1(t))^2$ is continuous function in $[0,T]$.

We now consider conditions for existence of a transformation $y = y(t)$ satisfying the condition (3.4.4) and how to construct the transformation.

THEOREM 3.4.1. *Let* $\Omega_0 = \{(x,t) : \sigma(x,t) = 0,\ (x,t) \in \Omega\}$, $\Omega_1(t) = \{x : \sigma(x,t) > 0,\ x \in (\mu_1(t),\mu_2(t))\}$, $\Omega_2(t) = \{x : \sigma(x,t) < 0,\ x \in (\mu_1(t),\mu_2(t))\}$,

$c(x,t)$ and $r(x,t)$ be continuous functions defined on Ω. For $t \in (0,T)$ define

$$\theta_1(t) = \begin{cases} \inf\limits_{x\in\Omega_1(t)} \dfrac{1}{\sigma}(c-\sigma_t-r), & \text{if } \Omega_1(t) \neq \emptyset, \\ \sup\limits_{x\in\Omega_2(t)} \dfrac{1}{\sigma}(c-\sigma_t-r), & \text{if } \Omega_1(t) = \emptyset, \end{cases}$$

$$\theta_2(t) = \begin{cases} \sup\limits_{x\in\Omega_2(t)} \dfrac{1}{\sigma}(c-\sigma_t-r), & \text{if } \Omega_2(t) \neq \emptyset, \\ \inf\limits_{x\in\Omega_1(t)} \dfrac{1}{\sigma}(c-\sigma_t-r), & \text{if } \Omega_2(t) = \emptyset. \end{cases}$$

If $\sigma_t(x,t)+r < c(x,t)$ for all $(x,t) \in \Omega_0$ and $\theta_1(t) \geq \theta_2(t)$, then there exist a smooth function $y = y(t)$ satisfying

$$\sigma_t + \sigma(\log y')' + r \leq c(x,t). \tag{3.4.5}$$

Proof. It is readily seen that $\Omega_1(t)$ and $\Omega_2(t)$ are open sets for fixed $t \in (0,T)$. Denote

$$\Omega_1 = \{t:\ \Omega_1(t) \neq \emptyset,\ t \in (0,T)\}, \qquad \Omega_2 = \{t:\ \Omega_2(t) \neq \emptyset,\ t \in (0,T)\}.$$

Then Ω_1 and Ω_2 are also open sets.

Since $\sigma_t(x,t)+r < c(x,t)$ for all $(x,t) \in \Omega_0$ and there is no $t \in (0,T)$ such that $\sigma(x,t) = 0$ for all $x \in (a,b)$, functions $\theta_1(t)$ and $\theta_2(t)$ are well defined for $t \in (0,T)$. On the other hand, it is straightforward to show that $\theta_1(t)$ and $\theta_2(t)$ are continuous on Ω_1 and Ω_2, respectively. This implies that there is no $t \in (0,T)$ such that both $\theta_1(t)$ and $\theta_2(t)$ are discontinuous at t because $\Omega_1 \cup \Omega_2 = (0,T)$. Hence, there exists a continuous function $\theta(t)$ such that $\theta_2(t) \leq \theta(t) \leq \theta_1(t)$. Let y be a solution of $(\log y')' = \theta(t)$. Then y is a smooth function on $(0,T)$. If $(x,t) \in \Omega_0$, it is clear that (3.4.5) holds. If $(x,t) \in \Omega$ and $x \in \Omega_1(t)$, a straightforward computation shows that

$$\sigma_t + \sigma(\log y')' + r \leq \sigma_t + \sigma\theta(t) + r \leq \sigma_t + r + \sigma\Big(\frac{1}{\sigma}(c-\sigma_t-r)\Big) = c.$$

If $(x,t) \in \Omega$ and $x \in \Omega_2(t)$, the conclusion follows in a similar way. □

THEOREM 3.4.2. *If $\sigma = \kappa(x)\varphi(t)$, $\varphi(t) \in C^1(0,T)$ and $\varphi(t) \neq 0$ for $t \in (0,T)$, then there exists a smooth function $y(t) \in [0,T]$ such that*

$$y(0) = 0, \quad y'(t) > 0,\ \forall t \in (0,T), \quad y(T) < +\infty, \tag{3.4.6}$$

$$\sigma_t + \sigma(\log y')' < c(t) < 2\pi^2/(\mu_2(t) - \mu_1(t))^2, \tag{3.4.7}$$

Proof. Without loss of generality, assume that $\varphi(t) > 0$ for $t \in (0,T)$. If there exist positive constants C_1, C_2 and sufficient small δ_1 and δ_2 such that

$$\varphi(t) \geq C_1 t, \quad t \in (0,\delta_1), \quad \varphi(t) \geq C_2(T-t), \quad t \in (T-\delta_2, T) \tag{3.4.8}$$

Let $c(x,t) = \varepsilon\kappa(x)\varphi(t)$, where ε is a positive constant such that $c(x,t) \leq c < 2\pi^2/(\mu_2(t) - \mu_1(t))^2$. Let $r(x,t) = 0$. It follows from Theorem 3.4.1 that

$$\theta_1(t) = \theta_2(t) = \theta(t) = (\varepsilon - 1)\varphi'(t)/\varphi(t).$$

Let

$$y = M\int_0^t \Big(\frac{1}{\varphi(t)}\Big)^{1-\varepsilon} dt.$$

Then y satisfies the condition (3.4.6). Theorem 3.4.1 shows that y satisfies (3.4.7).

Otherwise, we have $\varphi \to 0$ and $\varphi'(t) \to 0$ if $t \to 0$ or $t \to T$. Denote $\Omega_\delta = \{t;\ t \in (0,T)$ and $\varphi'(t) < \delta\}$, where δ is a positive constant such that $\kappa(x)\delta < 2\pi^2/(\mu_2(t) - \mu_1(t))^2$ for $(x,t) \in \Omega$ and $t \in (0,T) - \Omega_t$. Define

$$c(x,t) = \begin{cases} \kappa(x)\varphi'(t), & \text{if } (x,t) \in \Omega \text{ and } t \in \Omega_\delta \\ \kappa(x)\delta, & \text{if } (x,t)\Omega \text{ and } t \in (0,T) - \Omega_\delta \end{cases}$$

A simple computation shows that

$$\theta_1(t) = \theta_2(t) = \theta(t) = \begin{cases} 0, & \text{if } t \in \Omega_\delta \\ (\delta - \varphi'(t))/\varphi(t), & \text{if } t \in (0,T) - \Omega_\delta \end{cases}$$

Let

$$y = M \int_0^t \exp\Big(\int_0^t \theta(t)dt \Big) dt.$$

It is straightforward to show that (3.4.6) holds. Furthermore, Theorem 3.4.1 shows that y satisfies (3.4.7). □

If $2p - q_x \geq 0$ and $\sigma(x,t) = \kappa(x)\varphi(t)$, Theorem 3.4.2 shows that we can always find transformations $x = x$ and $y = y(t)$ such that (3.4.4) holds for $c_1(t) < 2\pi^2/(\mu_2(t) - \mu_1(t))^2$ because

$$\sigma_t + \sigma(\log y')' + \widetilde{q}_x - 2\widetilde{p} = \sigma_t + \sigma(\log y')' + q_x - 2p \leq \sigma_t + \sigma(\log y')'.$$

Therefore, we can use our method to solve the problem (3.4.1), (3.4.2). In particular, we can solve the model problem of $p = q = 0$ if $\sigma = \kappa(x)\varphi(t)$ without any other condition.

Now we consider the case where $q = 0$, $p \geq 0$ and σ is a function of $x + ct + d$, i.e., $\sigma = \sigma(x + ct + d)$, where c is a constant. With the use of the transformation $t = t$ and $y = x + ct + d$ we obtain a new equation

$$\sigma(y)v_t = v_{yy} - c\sigma(y)v_y - pv + f, \qquad \forall (y,t) \in \widetilde{\Omega}, \tag{3.4.9}$$

$$\begin{cases} v(\mu_1(t) + ct + d, t) = v(\mu_2(t) + ct + d, t) = 0, & \forall t \in (0,T), \\ v(y,0) = u_0(y - d) & \text{for } \sigma(y) > 0, \\ u(x,T) = u_T(y - cT - d) & \text{for } \sigma(y) < 0, \end{cases} \tag{3.4.10}$$

where $\widetilde{\Omega} = \{(y,t) : \mu_1(t) + ct + d < y < \mu_2(t) + ct + d,\ 0 < t < T\}$. Let function $w = \exp(-\frac{c}{2}\int_0^y \sigma(z)dz)v$. Then the problem (3.4.9), (3.4.10) becomes

$$\sigma(y)w_t = w_{yy} - \widetilde{p}w + \widetilde{f}, \qquad \forall (y,t) \in \widetilde{\Omega}, \tag{3.4.11}$$

$$\begin{cases} w(\mu_1(t) + ct + d, t) = w(\mu_2(t) + ct + d, t) = 0, & \forall t \in (0,T), \\ w(y,0) = w_0(y) & \text{for } \sigma(y) > 0, \\ w(x,T) = w_T(y) & \text{for } \sigma(y) < 0, \end{cases} \tag{3.4.12}$$

where

$$\widetilde{p} = p + \Big(\tfrac{c}{2}\sigma(y)\Big)^2 - \tfrac{c}{2}\sigma'(y), \qquad \widetilde{f} = \exp\Big(-\tfrac{c}{2}\int_0^y \sigma(z)dz\Big)f,$$

$$w_0(y) = \exp\Big(-\tfrac{c}{2}\int_0^y \sigma(z)dz\Big)u_0(y-d),$$

$$w_T(y) = \exp\Big(-\tfrac{c}{2}\int_0^y \sigma(z)dz\Big)u_T(y-cT-d).$$

We now prove that our method is applicable to the problem (3.4.11), (3.4.12). By using Theorem 3.2.1, our task is to search a smooth function $g(y)$ such that

$$g' + 2\widetilde{p} \geq g' + p - c\sigma' + \frac{c^2}{2}\sigma^2 \geq g' - c\sigma' + \frac{c^2}{2}\sigma^2 > \frac{1}{2}g^2,$$

i.e.,

$$r' > \frac{1}{2}r(r+2\sigma), \tag{3.4.13}$$

where $r = g - c\sigma$. Assume that $|c\sigma| < M$ for all $(y,t) \in \widetilde{\Omega}$ and $M \geq 1$. It is readily seen that $r = 2\varepsilon\exp(My)/(1-\varepsilon\exp(My))$ satisfies (3.4.13), where ε is a positive constant such that $1 - \varepsilon\exp(My) > 0$ for $(y,t) \in \widetilde{\Omega}$. Therefore $g = c\sigma + r$.

A example given in [11] shows that if $\sigma_t \geq \pi^2/2$ in every point of Ω directly using Lu's method fails to solve the forward-backward heat equation. The situation is the same for the generalization in the present paper simply because the generalization is reduced to Lu's method if $\mu_1(t) = -1$, $\mu_2(t) = 1$ and $p = q = u_0(x) = u_T(x) = 0$. However, we can solve all forward-backward heat equation for which $\sigma(x,t) = \sigma(x+ct+d)$, $q = 0$ and $p \geq 0$. Therefore, the use of a variable transformation leads indeed to a substantial improvement of the generalization of Lu's method presented in this paper.

3.5. Numerical Tests

Most test examples involve identical numerical methods which are commented upon below first. Afterwards, the numerical tests are presented. Emphasis is placed on examples out of the extended class of problems of the form (3.1.1) which can be solved with the use of the theory presented in this paper.

We now consider the numerical methods which are used to perform the tests. Except when mentioned otherwise the methods described below are used for all tests. First, consider the mesh generation and refinement. The coarse grids are of the Tucker-Whitney triangular type described by Todd in [14]. The grid refinement used to create finer (uniform) meshes is of the bisection type applicable to arbitrary space dimensions as described in Maubach [13]. Secondly, the finite element bases used are conforming bases of polynomial degree k which are for instance described by Zienkiewicz in [22]. Whenever the polynomial degree is not mentioned explicitly, degree $k = 1$ is assumed.

The discrete systems were all solved with the use of iterative solution methods. Because the global finite element presented in this paper to solve the forward-backward heat equation always leads to a non-symmetric discrete system due to the discretization of the time variable, the Preconditioned Conjugate Gradient method (PCG) was not considered. Here, all discrete solutions are computed with the Conjugate Gradient Squared (CGS) by Sonneveld from [17] and in all cases this method is accelerated with an ILU(0) preconditioner. Also the Generalized Minimal

Residual method (GMRES) by Saad and Schultz described in [16], the Generalized Conjugate Gradient Least Squares method (GCGLS) by Axelsson in [2] and [1] were used for several tests. As they took usually few more iterations, all tests are presented with CGS for sake of brevity and to make comparison easier. The initial solution used by the iterative solver is a vector of which each coefficient is related to one finite element basis function and corresponding support point, called node. Below, this vector has the correct value for nodes at the Dirichlet boundary, and the value 0 at all other nodes. The stopping criterion is of the form

$$||r^{(k)}||_2 < \epsilon \tag{3.5.1}$$

where, unless mentioned otherwise, $\epsilon = 10^{-8}$. Here, $r^{(k)}$ is the k-th updated residual (as is usual with many iterative methods, CGS updates the residual independently from the solution in order to save a matrix multiplication) which is identical to $Ax^{(k)} - b$ if there are no round-off errors. The difference between $r^{(k)}$ and $Ax^{(k)} - b$ was monitored and turned out to be negligible for all tests. Finally, the ILU(0) preconditioner (due to its inexact nature) depends on the numbering of the finite element basis functions. For uniform meshes and diffusion dominated problems, the preconditioner works well independent on the numbering but for locally refined meshes this is not necessarily the case. In all examples, the nodes were numbered left right and bottom up. For a few examples a different numbering was tested (right to left, top to bottom) but no large discrepancies in numbers of iterations were found. As usual, the ILU(0) preconditioner performs well for piecewise linear $k = 1$ finite element discretization of (3.1.1) when $q = 0$ and $p > 0$.

The discrete system's coefficients were computed with quadrature were as usual the degree of the rule is taken to be $2k$, in order to take into account the term cuv in the variational formulation. However, in the case where the solution or one of its derivatives is non-continuous (see below) across an element edge a quadrature rule was used with only quadrature points strictly interior to the element. In the case of $k = 1$, for instance, we used rule "T_2: 5-1" of degree 5 from Stroud [20], page 314. For higher k, most quadrature formulas were taken from the paper [5] by Duvenant. The integrals over the edges involved with the $||\,.\,||_{(1,0)}$ norm were computed with the standard Gauss-Legendre formulas of degree $2k$.

Finally, the error estimates shown in the tables below were obtained by quadrature identical to the quadrature used for the construction of the linear systems of equations, i.e., we made no use of quadrature rules of order higher than $2k$. Each table with measured convergence rates will show per row, from the first column to the last one in order: the minimum element diameter of the related mesh, the number of unknowns for the this mesh and the calculated rate of convergence of $u - u^h$ measured with an approximate L^∞ norm, L^2 norm, H^1 semi-norm

$$||u|| = \left(\int_0^1 \int_\Omega u_x^2 + u_t^2 \, dx \, dt \right)^{1/2} \tag{3.5.2}$$

and the $||\,.\,||_{(1,0)}$ norm, abbreviated with W. Due to the time-derivative in this semi H^1 norm, the W norm errors may be less than errors in this former norm. Tables which contain the actual error will show the actual errors in the same order and do not contain the column which shows the number of unknowns.

It is reconsidered in order to corroborate the error estimate (3.3.2) in Theorem 3.3.2 for several degrees k of finite element polynomial approximations. The

equation is

$$\sigma u_t = u_{xx} - q u_x - p u + f, \quad \forall (x,t) \in \Omega = (-1,1) \times (0,1), \tag{3.5.3}$$

complemented with boundary conditions as in (3.1.2). For the first example, we choose $p = q = 0$ and $\sigma = (\pi^2 - 1)x/2$. The function f is taken such that the solution of the above equation is given by

$$u(x,t) = \begin{cases} (x-1)^2 t^2 [(t-1)^2 - 4x^2] & \forall x \geq 0,\ t \in [0,1], \\ (x-1)^2 (t-1)^2 (t^2 - 4x^2) & \forall x < 0,\ t \in [0,1]. \end{cases}$$

The solution has a discontinuous second derivative in x, whence a higher order quadrature rule is used as described above. Because $q = p = 0$ and $\sigma_t = 0$, both Corollary 3.2.2 and Corollary 3.2.3 in combination with Theorem 3.2.1 show that the solution to (3.5.3) is unique.

Tables 1 and 3 below show the rate of convergence of $u - u^h$ on subsequent uniform grids ($h_{i+1} = h_i/2$) (approximates the exponents k in (3.3.2)) for degrees 1 and 3 polynomial finite element approximation, respectively. On the finest mesh the discrete Galerkin systems resulting from (3.5.3) involves 33153 and 74305 unknowns. Tables 2 and 4 shows the size of the actual errors.

TABLE 1. Galerkin convergence rates for example 3, $k = 1$.

h	N	max $\lvert e \rvert$	L^2	H^1	W
0.250000	45	1.263	2.159	0.823	1.003
0.125000	153	1.595	2.252	1.056	1.159
0.062500	561	1.800	2.094	1.023	1.062
0.031250	2145	1.894	2.033	1.008	1.017
0.015625	8385	1.929	2.015	1.003	1.004
0.007812	33153	1.955	2.008	1.001	1.001

TABLE 2. Errors of Galerkin method for example 3, $k = 1$.

h	max $\lvert e \rvert$	L^2	H^1	W
0.500000	0.68D+00	0.47D+00	0.23D+01	0.17D+01
0.250000	0.28D+00	0.10D+00	0.13D+01	0.88D+00
0.125000	0.94D-01	0.22D-01	0.65D+00	0.39D+00
0.062500	0.27D-01	0.52D-02	0.32D+00	0.19D+00
0.031250	0.72D-02	0.12D-02	0.16D+00	0.94D-01
0.015625	0.19D-02	0.31D-03	0.79D-01	0.46D-01
0.007812	0.49D-03	0.79D-04	0.39D-01	0.23D-01

TABLE 3. Galerkin convergence rates for example 3, $k = 3$.

h	N	max $\|e\|$	L^2	H^1	W
0.250000	325	4.155	4.250	3.419	3.082
0.125000	1225	3.839	4.062	2.969	3.031
0.062500	4753	3.683	4.021	2.986	3.001
0.031250	18721	3.804	4.002	2.993	2.995
0.015625	74305	3.880	3.992	2.992	2.996

TABLE 4. Errors of Galerkin method for example 3, $k = 3$.

h	max $\|e\|$	L^2	H^1	W
0.500000	0.31D-01	0.12D-01	0.23D+00	0.94D-01
0.250000	0.17D-02	0.64D-03	0.21D-01	0.11D-01
0.125000	0.12D-03	0.38D-04	0.27D-02	0.13D-02
0.062500	0.96D-05	0.23D-05	0.34D-03	0.17D-03
0.031250	0.69D-06	0.14D-06	0.43D-04	0.21D-04
0.015625	0.46D-07	0.93D-08	0.54D-05	0.26D-05

For the second case, the solution u to equation (3.5.3) is identical to the previous solution, but here we take $p = -4x$, $p = -4$ and adapt f. For the sake of simplicity, σ is taken to be the same as in the first example. According to Corollary 3.2.2, $\sigma_t + q_x - 2p = 4$ whence the requirement (3.2.2) is satisfied and a unique solution u exists according to Theorem 3.2.1. Tables 5 through 8 show the rates of convergence and actual errors for a $k = 1$ and $k = 3$ solution. Both tests corroborate to the predicted $||\,.\,||_{(1,0)}$-convergence rate of order k by Theorem 3.3.2. The observed rate of convergence in the L^2 norm, not theoretically founded in this paper, turns out to be of the order $k + 1$.

TABLE 5. Galerkin convergence rates for example 4, $k = 1$.

h	max $\|e\|$	L^2	H^1	W
0.250000	1.922	2.007	0.951	1.090
0.125000	1.618	2.295	1.085	1.187
0.062500	1.833	2.118	1.035	1.070
0.031250	1.908	2.044	1.013	1.019
0.015625	1.943	2.023	1.006	1.005
0.007812	1.966	2.014	1.003	1.001

TABLE 6. Errors of Galerkin method for example 4, $k = 1$.

h	max $\|e\|$	L^2	H^1	W
0.500000	0.11D+01	0.68D+00	0.31D+01	0.19D+01
0.250000	0.30D+00	0.17D+00	0.16D+01	0.91D+00
0.125000	0.10D+00	0.34D-01	0.77D+00	0.40D+00
0.062500	0.28D-01	0.80D-02	0.37D+00	0.19D+00
0.031250	0.74D-02	0.19D-02	0.18D+00	0.94D-01
0.015625	0.19D-02	0.47D-03	0.93D-01	0.46D-01
0.007812	0.49D-03	0.11D-03	0.46D-01	0.23D-01

TABLE 7. Galerkin convergence rates for example 4, $k = 3$.

h	max $\|e\|$	L^2	H^1	W
0.250000	4.597	4.005	3.561	3.096
0.125000	3.895	4.007	2.961	3.038
0.062500	3.722	4.019	2.978	3.003
0.031250	3.824	4.042	2.984	2.996

TABLE 8. Errors of Galerkin method for example 4, $k = 3$.

h	max $\|e\|$	L^2	H^1	W
0.500000	0.47D-01	0.16D-01	0.26D+00	0.95D-01
0.250000	0.19D-02	0.99D-03	0.22D-01	0.11D-01
0.125000	0.13D-03	0.61D-04	0.29D-02	0.13D-02
0.062500	0.99D-05	0.38D-05	0.37D-03	0.17D-03
0.031250	0.70D-06	0.23D-06	0.46D-04	0.21D-04

Here, due to a negative mass term and higher polynomial degree ($k = 3$) the ILU(0) preconditioner performed rather poorly. Orthogonality of the iterative solution method (here CGS) is partially lost: the first time the updated residual met the stop criterion its magnitude was $\|r^{(k)}\|_2 = 0.992d - 11$ while the actual residual turned out to be $\|Ax^{(k)} - b\|_2 = 0.617d - 05$, several orders of magnitude larger. This was overcome with a restart of CGS taking the newly computed approximate solution $x^{(k)}$ to be the initial solution (such strategy would not work for stationary iterative solution methods).

Finally, consider example 3 where one takes $\sigma = acx(t+1)$ with $a = 2$ and $c = \pi^2/2$ and where $p = q = 0$. Because σ is separable, we can apply the transformation $y(t) = \ln(t+1)/\ln(2)$.

The transformed problem is solved on subsequent uniform grids, as previously created with refinement, where the mesh size parameter is given by $h = 1/2^i$, $i \geq 2$. After the discrete solution to the transformed equation (but on the uniform grid) is computed, the grid and its related solution is backwards transformed via $t = 2^y - 1$, after which table 10 with discretization errors and 9 with rates of convergence are computed. Clearly, also under this transformation the discretization error estimate holds.

TABLE 9. Galerkin convergence rates for example 5, $k = 1$.

h	max $\|e\|$	L^2	H^1	W
0.250000	2.037	2.681	1.073	1.276
0.125000	1.599	2.567	1.230	1.378
0.062500	1.646	2.241	1.106	1.159
0.031250	1.848	2.071	1.029	1.045
0.015625	1.942	2.020	1.007	1.011
0.007812	1.934	2.006	1.002	1.003

TABLE 10. Errors of Galerkin method for example 5, $k = 1$.

h	max $\|e\|$	L^2	H^1	W
0.500000	0.96D+00	0.64D+00	0.28D+01	0.28D+01
0.250000	0.23D+00	0.10D+00	0.13D+01	0.11D+01
0.125000	0.77D-01	0.17D-01	0.58D+00	0.44D+00
0.062500	0.24D-01	0.36D-02	0.27D+00	0.20D+00
0.031250	0.68D-02	0.85D-03	0.13D+00	0.96D-01
0.015625	0.17D-02	0.21D-03	0.66D-01	0.48D-01
0.007812	0.46D-03	0.52D-04	0.33D-01	0.24D-01

TABLE 11. Galerkin convergence rates for example 5, $k = 3$.

h	max $\|e\|$	L^2	H^1	W
0.250000	3.786	4.246	3.195	3.138
0.125000	3.808	4.175	2.975	3.066
0.062500	3.952	4.064	2.978	3.020
0.031250	3.703	4.018	2.994	3.002

TABLE 12. Errors of Galerkin method for example 5, $k = 3$.

h	max $\|e\|$	L^2	H^1	W
0.500000	0.23D-01	0.10D-01	0.15D+00	0.10D+00
0.250000	0.17D-02	0.53D-03	0.17D-01	0.11D-01
0.125000	0 12D-03	0.29D-04	0.22D-02	0.13D-02
0.062500	0 79D-05	0.17D-05	0.28D-03	0.17D-03
0.031250	0 60D-06	0.11D-06	0.35D-04	0.21D-04

References

[1] O. AXELSSON, *Conjugate gradient type methods for unsymmetric and inconsistent systems of linear equations*, Linear Algebra Appl., 29 (1980), pp. 1–16.

[2] ———, *A restarted version of a generalized preconditioned conjugate gradient method*, Communications in Applied Numerical Methods, 4 (1988), pp. 521–530.

[3] A. K. AZIZ AND J.-L. LIU, *A Galerkin method for the forward-backward heat equation*, Math. Comp., 56 (1991), pp. 35–44.

[4] M. S. BAOUENDI AND P. GRISVARD, *Sur une équation d'évolution changegeant de type*, J. Funct. anal., 2 (1968), pp 352–367.

[5] DUVENANT, *High degree efficient symmetric gauss quadrature rules for the triangle*, International Journal for Numerical Methods in Engineering, 21 (1985), pp. 11291–11486.

[6] J. A. FRANKLIN AND E. R. RODEMICH, *Numerical analysis of an elliptic-parabolic partial differential equation*, SIAM J. Numer. Anal., 5 (1968), pp. 680–716.

[7] M. GEVREY, *Sur les équations aux dérivées partielles du type parabolique*, J. Math. pures Appl., 6 (1913), pp. 305–475.

[8] ———, *Sur les équations aux dérivées partielles du type parabolique (suite)*, J. Math. pures Appl., 6 (1914), pp. 105–148.

[9] T. LAROSA, *The propagation of an electron beam through the solar corona*, PhD thesis, Department of Physics and Astronomy, University of Maryland, 1986.

[10] J. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod Gauthier-Villars, Paris, 1969.

[11] H. LU, *Galerkin and weighted Galerkin methods for the forward-backward heat equation*, tech. rep., 9410, Department of Mathematics, University of Nijmegen, The Netherlands, 1994.

[12] H. LU AND Z.-Y. WEN, *Solution of a forward-backward heat equation*, Manuscript, (1994).

[13] J. MAUBACH, *Local bisection refinement for n-simplicial grids generated by reflections*, SIAM J. Sci. Comput, 16 (1995).

[14] J. T. MICHAEL, *The Computation of Fixed Points and Applications, Lecture Notes in Economics and Mathematical Systems 124*, Springer Verlag, Berlin, 1967.

[15] J. NEČAS, *Les Méhtodes Directes en Théorie des Equations Elliptiques*, Masson, Paris, 1967.

[16] Y. SAAD AND M. H. SCHULTZ, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput, 7 (1986), pp. 856–869.

[17] P. SONNEVELD, *CGS, a fast Lanczos-type solver for non-symmetric linear systems*, SIAM J. Sci. Stat. Comput, 10 (1989), pp. 36–52.

[18] K. STEWARTSON, *Multistructural boundary layers on flat plates and related bodies*, Adv. in Appl. Mech., 14 (1974), pp. 145–239.

[19] ———, *D'Alembert's paradox*, SIAM Rev., 23 (1981), pp. 308–343.

[20] A. H. STROUD, *Approximate Calculation of Multiple Integrals*, Prentice-Hall (Series in Automatic Computation), New York, 1971.

[21] V. VANAJA AND R. B. KELLOGG, *Iterative methods for a forward-backward heat equation*, SIAM J. Numer. Anal., 27 (1990), pp. 622–635.

[22] O. ZIENKIEWICZ, *The Finite Element Method in Engineering Science*, 3rd *edition*, Mc Graw-Hill, New York, 1977.

CHAPTER 4

A Barrier on Finite-Difference Schemes of Positive Type*

Abstract. In this note it is shown a uniform consistency barrier on finite difference schemes of positive type for convection-diffusion equations, i.e., any difference scheme of positive type cannot approximate $Lu = -\varepsilon\Delta u + \vec{f}\cdot\nabla u + gu$ to $O(h^\alpha)$ $(\alpha > 1)$ accuracy uniformly in ε.

Key words. difference schemes of positive type, truncation error, uniform-difference scheme, convection-diffusion equation

AMS subject classifications. 65L05, 65N06

4.1. Introduction

Consider a discrete method for solving a differential equation $Lu = f$. Let L_h be a discrete approximation to L defined on a difference or finite element mesh Ω_h depending on some meshwidth h. L_h is monotone if $L_h v \geq 0$ implies $v \geq 0$, where v is a function defined on Ω_h. Let w be a barrier function, i.e., a normalized function with $\max w(x_i) = 1$, such that $L_h w(x_i) \geq c > 0$, $\forall x_i \in \Omega_h$. If L_h is monotone, it is shown (see [1] for instance) that

$$\|L_h^{-1}\|_\infty \leq c^{-1},$$

where $\|L_h\|_\infty = \sup_{v\neq 0} \|L_h v\|_\infty / \|v\|_\infty$ and $\|v\|_\infty = \max_{x_i\in\Omega_h} |v(x_i)|$ is the maximum norm, and the discretization error $\|u - u_h\|_\infty$ is bounded by the truncation error $\|L_h u - f\|_\infty$ as follows:

$$\|u - u_h\|_\infty \leq c^{-1}\|L_h u - f\|_\infty.$$

In general, it is not easy to check if a discrete difference linear operator L_h is monotone. Consider a finite difference scheme

$$L_k^h u = \Big(2a_0 \;\; {}_0T(0,\ldots,0) - \sum_{i_1=-p_1}^{q_1} \cdots \sum_{i_k=-p_k}^{q_k} a_{i_1} \;\; {}_{i_k}T(i_1,\ldots,i_k)\Big)u \tag{4.1.1}$$

in k-D for a uniform mesh, where u is a function defined on Ω_h, $T(\beta_1,\beta_2,\ldots,\beta_k)$ is a translation operator defined by

$$T(\beta_1,\beta_2,\ldots,\beta_k)u(x_1,x_2,\ldots,x_k) = u(x_1+\beta_1 h, x_2+\beta_2 h,\ldots,x_k+\beta_k h),$$

* This chapter is based on the paper: H. Lu, *A uniform-consistency barrier on finite-difference schemes of positive type for convection-diffusion equations*, SIAM J. Sci. Comput. 16 (1995), pp 169-172.

and the β_i's are integers. L_k^h is of positive type if

$$a_{i_1 i_2 \ldots i_k} \geq 0, \tag{4.1.2}$$
$$i_1 = -p_1, \ldots, q_1,\ i_2 = -p_2, \ldots, q_2, \ \ldots, \ i_k = -p_k, \ldots, q_k,$$

$$2a_{0\ 0} \geq \sum_{i_1=-p_1}^{q_1} \sum_{i_2=-p_2}^{q_2} \cdots \sum_{i_k=-p_k}^{q_k} a_{i_1 i_2\ \ldots i_k}. \tag{4.1.3}$$

It is well known that L_k^h is automatically monotone if L_k^h is of positive type. Hence much attention has been paid to difference schemes of positive type [3], [5], [7], [8], [9]. In particular, difference schemes are useful for numerical methods for singular perturbation problems [3], [4], [5], [6], [7], [8], [9]. Unfortunately, in 1978, Kellogg and Tsan [7] showed that any 3-point difference scheme of positive type cannot approximate $L(u) = -\varepsilon u'' + b(x)u' + g(x)u$ to $O(h^\alpha)$ $(\alpha > 1)$ accuracy uniformly in ε. This implies that it is difficult to obtain a highly accurate approximation of the solution if any 3-point difference scheme of positive type with meshpoints independing on ε is used to solve a convection-diffusion equation $-\varepsilon u'' + b(x)u' + g(x)u = 0$. Note that this barrier can be overcome somehow in one dimension by slightly adding meshpoints depending on ε. Recently, Axelsson and Nikolova showed an $O(h^2)$ accuracy uniformly in ε by using $O(\log \varepsilon^{-1} + h^{-1})$ points [2].

The aim of this note is to show that any difference scheme of positive type cannot approximate $Lu = -\varepsilon \Delta u + \vec{f} \cdot \nabla u + gu$ to $O(h^\alpha)$ $(\alpha > 1)$ accuracy uniformly in ε. It is shown first that any difference scheme of positive type cannot approximate $L(u) = -\varepsilon u'' + b(x)u' + g(x)u$ to $O(h^\alpha)$ $(\alpha > 1)$ accuracy uniformly in ε, which generalizes the result given by Kellogg and Tsan [7]. We then extend the result to higher dimensions.

4.2. Main Results

Let L_h be a discrete approximation to an operator L defined on a difference mesh Ω_h depending on some meshwidth h and $\sigma(h)$ be a positive function of the meshwidth h. If there exists a positive constant C independent of h such that

$$|L_h v(x_i) - Lv(x_i)| \leq C\sigma(h), \forall x_i \in \Omega_h, \tag{4.2.1}$$

where v is a smooth function, it is said that L_h approximates L to $O(\sigma(h))$ accuracy.

Let

$$L^h v_n = a_0(h,n)v_n - \sum_{i=-p}^{-1} a_i(h,n)v_{n+i} - \sum_{i=1}^{q} a_i(h,n)v_{n+i} \tag{4.2.2}$$

denote an approximation to the operator $Lu = -\varepsilon u'' + f(x)u' + g(x)u$, where p and q are nonnegative integers and $f(x)$ is not identically zero. First we prove that any difference scheme L^h of positive type cannot approximate L to $O(h^\alpha)$ $(\alpha > 1)$ accuracy uniformly in ε.

THEOREM 4.2.1. *Suppose that L^h approximates L to $O(h^\alpha)$ $(\alpha > 1)$ accuracy uniformly in ε. Then L^h is not of positive type.*

Proof. Denote $a_i = a_i(h,n)$, $i = -p, \ldots, q$, for convenience in the proof. Under the assumption of the theorem, $L^h(x^k) = L(x^k) + \gamma_k$ for $k = 0, 1, 2$, where

$|\gamma_k| \le C_k h^\alpha$, C_k is independent of h and ε. Hence

$$
\begin{aligned}
&-\sum_{i=-p}^{-1} a_i + a_0 - \sum_{i=1}^{q} a_i = g(x) + \gamma_0, \\
&-\sum_{i=-p}^{-1} a_i(x+ih) + a_0 x - \sum_{i=1}^{q} a_i(x+ih) = f(x) + g(x)x + \gamma_1, \\
&-\sum_{i=-p}^{-1} a_i(x+ih)^2 + a_0 x^2 - \sum_{i=1}^{q} a_i(x+ih)^2 \\
&= -2\varepsilon + 2f(x)x + g(x)x^2 + \gamma_2.
\end{aligned}
$$

By direct computation, one obtains the following equations.

$$
\begin{aligned}
&-\sum_{i=-p}^{-1} a_i + a_0 - \sum_{i=1}^{q} a_i = g(x) + \gamma_0, \\
&\sum_{i=-p}^{-1} i a_i - \sum_{i=1}^{q} i a_i = (f(x) + \gamma_1 - x\gamma_0)h^{-1}, \\
&-\sum_{i=-p}^{-1} i^2 a_i - \sum_{i=1}^{q} i^2 a_i = (-2\varepsilon + \gamma_2 - 2x\gamma_1 + x^2\gamma_0)h^{-2}.
\end{aligned}
$$

Adding the last two equations shows

$$
\begin{aligned}
(4.2.3) \qquad &\sum_{i=1}^{p} i(i-1)a_{-i} + \sum_{i=1}^{q} i(i+1)a_i \\
= \; &(2\varepsilon - \gamma_2 + 2x\gamma_1 - x^2\gamma_0)h^{-2} - (f(x) + \gamma_1 - x\gamma_0)h^{-1}.
\end{aligned}
$$

Since $f(x)$ is not identically zero, one can find a point x such that either $f(x) > 0$ or $f(x) < 0$. In the former case, since $|\gamma_k| \le C_k h^\alpha$, $k = 0, 1, 2$, (4 2.3) becomes strictly negative for ε sufficiently small, which implies L^h is not of positive type. In the latter case, doing a transformation $x \mapsto c - x$, we have similarly that

$$
\begin{aligned}
(4.2.4) \qquad &\sum_{i=1}^{p} i(i-1)a_{-i} + \sum_{i=1}^{q} i(i+1)a_i \\
&= (2\varepsilon - \gamma_2 + 2(c-x)\gamma_1 - (c-x)^2\gamma_0)h^{-2} + (f(c-x) - \gamma_1 + (c-x)\gamma_0)h^{-1}.
\end{aligned}
$$

The conclusion follows from the same argument. □

In fact, we can see from the proof that L^h cannot approximate L to $O(h^\alpha)$ $(\alpha > 1)$ accuracy uniformly in ε if $a_i \ge 0$, $i = -p, \ldots, -1, 1, \ldots, q$. The result given by Kellogg and Tsan [7] is the case of $p = q = 1$.

Now we consider finite difference schemes in k-D. Let L_k^h defined by (4.1.1) denote an approximation to the operator

$$
(4.2.5) \qquad L_k u = -\varepsilon \Delta u + \vec{f} \cdot \nabla u + gu
$$

in a difference mesh Ω_h with a uniform meshwidth h, where vector-valued function $\vec{f} = (f_1(x_1), \ldots, f_k(x_k))^T$ is not identically zero. Since an ODE is a special case of a PDE, if something cannot be done for ODEs, it cannot in general be done

for PDEs. We claim that the same result of Theorem 4.2.1 holds in k-D. In fact, assume that $f_1(x_1)$ is not identically zero without loss of generality and let

$$c_0 = 2a_{0\ \cdots\ 0} - \sum_{i_2=-p_2}^{q_2} \cdots \sum_{i_k=-p_k}^{q_k} a_{0 i_2\ \cdots\ i_k}, \tag{4.2.6}$$

$$c_i = \sum_{i_2=-p_2}^{q_2} \cdots \sum_{i_k=-p_k}^{q_k} a_{i i_2\ \cdots\ i_k}, \tag{4.2.7}$$

$$i = -p_1, \ldots, -1, 1, \ldots, q_1.$$

If L_k^h is of positive type, one can see that

$$c_i \geq 0, \quad i = -p_1, \ldots, q_1, \quad 2c_0 \geq \sum_{-p_i}^{q_1} c_i. \tag{4.2.8}$$

Consider $L_k^h(x_1^j) = L_k(x_1^j) + \gamma_j'$, $j = 0, 1, 2$. The proof of Theorem 4.2.1 shows the following theorem.

THEOREM 4.2.2. *If L_k^h approximates L_k to $O(h^\alpha)$ $(\alpha > 1)$ accuracy uniformly in ε, then L_k^h is not of positive type.*

Theorem 4.2.2 reveals a uniform consistency barrier on finite difference schemes of positive type for convection-diffusion equations.

Acknowledgments

I am grateful to Professor W. Layton for suggesting the problem and valuable comments, and to Professor O Axelsson for valuable comments. I am also grateful to anonymous referees for helpful suggestions

References

[1] O. AXELSSON AND L. KOLOTILINA, *Monotonicity and discretization error estimates*, SIAM J. Numer. Anal., 27 (1990), pp. 1591–1611.

[2] O. AXELSSON AND M. NIKOLOVA, *Adaptive refinement for convection-diffusion problems based on defect-correction technique and finite difference methods*, in progress, (1995).

[3] V. ERVIN AND W. LAYTON, *A second order accurate, positive scheme for singularly perturbed boundary value problems*, Comp. Mech., 3 (1988), pp. 115–138.

[4] E. C. GARTLAND, JR, *Uniform high order difference schemes for singularly perturbed, two point, boundary value problems*, Math. Comp., 48 (1987), pp. 551–564.

[5] A. M. IL'IN, *Differencing scheme for a differential equation with a small parameter affecting the highest derivative*, Math. Notes, S S.R., 6 (1969), pp. 596–602.

[6] R. B. KELLOGG, *Analysis of a difference approximation for a singularly perturbed problem in two dimensions*, in Proc. B.A I.L., Dublin, (1980), pp. 113–117.

[7] R. B. KELLOGG AND A. TSAN, *Analysis of some difference approximations for a singular perturbation problem without turning points*, Math. Comp., (1978), pp 1025–1039.

[8] M. VELDHUIZEN, *Higher order methods for a singularly perturbed problem*, Numer. Math., 30 (1978), pp. 267–279.

[9] ———, *Higher order schemes of positive type for singular perturbation problems*, in Numer. and of Sing. Pert., P. W. Hemker and J. Miller, eds., Academic Press, New York, 1979, pp. 361–383.

Part 2

Analysis of Iterative Methods

CHAPTER 5

Numerical Radius and Application to Iterative Methods*

Abstract. Uses of the numerical radius in the analysis of basic iterative solution methods, of the SOR method for quasi-Hermitian positive definite matrices (not being consistently ordered) and of maximal eigenvalues of symmetric positive definite matrices using incomplete block-matrix factorizations are presented.

Key words. numerical radius, basic iterative method, SOR method, quasi-Hermitian positive definite matrix, incomplete block-matrix factorization, eigenvalue estimate

AMS subject classifications. 65F10, 65F15, 65F50

5.1. Introduction

Three uses of the numerical radius in the analysis of iterative solution methods are presented, where the use of the spectral radius is not applicable or would give inferior results.

When analyzing the rate of convergence of iterative solution methods commonly the spectral radius is used. However, for nonsymmetric iteration matrix B this gives only information about the asymptotic rate of convergence. It is shown that the numerical radius $r(B)$ is a more reliable measure of the convergence behavior for the initial iterations, because

$$r(B^m)^{\frac{1}{m}} \leq \|B^m\|^{\frac{1}{m}} \leq 2^{\frac{1}{m}} r(B^m)^{\frac{1}{m}} \leq 2^{\frac{1}{m}} r(B).$$

A matrix A is called a quasi-Hermitian positive definite if there exists a non-singular block diagonal matrix P such that PAP^{-1} is Hermitian positive definite. Next in the analysis of the successive overrelaxation method for quasi-Hermitian positive definite matrices it is shown that a crucial parameter (γ) depends on the numerical radius of the (block) lower triangular part ($\tilde{L}$) of the standard splitting of the matrix. On the other hand, in the analysis of the symmetric SOR method the corresponding parameter depends on the spectral radius of $\tilde{L}\tilde{L}^T$ (see Young [19], 1971 and Axelsson [3], 1974).

Finally, it is shown, using the numerical radius, that one can derive an upper bound of the largest eigenvalue of the preconditioned matrix $C^{-1}A$, when A is a symmetric and positive definite matrix partitioned in $m \times m$ blocks and C is an incomplete block factorization of A. Under a certain condition this upper bound is $2m$.

---

* This chapter is based on the paper, O. Axelsson, H. Lu and B. Polman, *On the numerical radius of matrices and its application to iterative methods*, Linear and Multilinear Algebra, 37(1994), pp. 225–238.

The paper is organized as follows: First in §5.2, some general results on the numerical radius and its relation with the spectral radius are presented. §5.3 presents estimates of the rate of convergence of basic iterative methods. In §5.4 the use of the numerical radius in the estimate of the rate of convergence of the SOR method for quasi-Hermitian positive definite matrices which are, in general, not consistently ordered, are found. Finally, in the last section the numerical radius is used for the derivation of an upper bound of the largest eigenvalue of a symmetric positive definite matrix preconditioned by a block incomplete factorization method. $\|\cdot\|$ denotes the 2-norm throughout the paper.

5.2. Numerical Radius of Matrices

Let A be an $n \times n$ complex matrix. The Rayleigh quotient of A for a vector $\mathbf{x} \neq 0$ is

$$q(\mathbf{x}) = (\mathbf{x}^* A\mathbf{x})/(\mathbf{x}^*\mathbf{x})$$

and the numerical radius of A is defined by

$$r(A) = \sup\{|\mathbf{x}^* A\mathbf{x}|;\ \mathbf{x} \in \mathbb{C}^n,\ \mathbf{x}^*\mathbf{x} = 1\}.$$

$V(A) = \{\mathbf{x}^* A\mathbf{x};\ \mathbf{x} \in \mathbb{C}^n,\ \mathbf{x}^*\mathbf{x} = 1\}$ is called the field of values or the numerical range of A.

It is readily seen that $r(A)$ is a matrix norm, i.e.,

1) $r(A) = 0$ if and only if $A = 0$
2) $r(\alpha A) = |\alpha| r(A)$, for any scalar $\alpha \in \mathbb{C}$
3) $r(A + B) \leq r(A) + r(B)$

but one important feature, namely multiplicativity, does not hold (see e.g. [7]). However, Pearcy [14] (see also Goldberg and Tadmor [7], Horn and Johnson [9]) proved the following power inequality

$$r(A^m) \leq r^m(A), \tag{5.2.1}$$

where m is a nonnegative integer. The norm $\|A\|$ is bounded by the numerical radius of A as follows (see e.g. Goldberg and Tadmor [7]):

$$(A) \leq \|A\| \leq 2r(A). \tag{5.2.2}$$

Based on inequality (5.2.1), we have the following result concerning the inverse of matrices.

THEOREM 5.2.1. *Let A be an $n \times n$ matrix and α be a constant such that $|\alpha| > r(A)$. Then $\alpha I - A$ is nonsingular and*

$$r((\alpha I - A)^{-1}) \leq \frac{1}{|\alpha| - r(A)}.$$

Proof. Let $(\lambda, \mathbf{x})$ be an eigenpair of A such that $\rho(A) = |\lambda|$ and $\|\mathbf{x}\| = 1$, then

$$\rho(A) = |\lambda| = |\mathbf{x}^* A\mathbf{x}| \leq r(A),$$

which implies $\alpha I - A = \alpha(I - \alpha^{-1}A)$ is nonsingular and

$$\rho(\alpha^{-1}A) \leq |\alpha^{-1}|\rho(A) \leq |\alpha^{-1}| r(A) < 1.$$

(5.2.1) shows that

$$\begin{aligned} r((\alpha I - A)^{-1}) &= |\alpha^{-1}| r((I - \alpha^{-1}A)^{-1}) \\ &= |\alpha^{-1}| r(\sum_{k=0}^{\infty} (\alpha^{-1}A)^k) \leq |\alpha^{-1}| \sum_{k=0}^{\infty} r^k(\alpha^{-1}A) \\ &\leq \frac{1}{|\alpha| - r(A)}, \end{aligned}$$

which leads the desired result. □

Denote by $H(A) = \frac{1}{2}(A + A^*)$ and $S(A) = \frac{1}{2}(A - A^*)$ the Hermitian and anti-Hermitian parts of A respectively. In 1975, Goldberg, Tadmor and Zwas [8] showed that for $A \geq 0$:

$$r(A) = \rho(H(A)) = \max_{\substack{\mathbf{x}^T\mathbf{x} = 1 \\ \mathbf{x} \in \mathbb{R}^n}} \mathbf{x}^T H(A)\mathbf{x}, \tag{5.2.3}$$

where $A \geq 0$ means that A is entry-wise nonnegative.

Let $M = \lambda I - A$. Then $(\mu, \mathbf{x})$ is an eigenpair of A if and only if $(\lambda - \mu, \mathbf{x})$ is an eigenpair of M. If $\rho(A) < |\lambda|$, it follows from the proof of Theorem 5.2.1 that M is nonsingular, and furthermore

$$\rho(M^{-1}) \leq \frac{1}{|\lambda| - \rho(A)}.$$

If, in addition, $A \geq 0$ and $\lambda > \rho(A)$,

$$M^{-1} = \lambda^{-1}(I - \lambda A)^{-1} = \lambda^{-1} \sum_{k=0}^{\infty} (\lambda^{-1}A)^k \geq 0$$

i.e., M is an M-matrix. However, it is sometimes difficult to compute $\rho(A)$ when A is nonsymmetric. Since $\rho(A) \leq r(A) = \rho(H(A))$ if $A \geq 0$, we can estimate the spectral radius of the inverse of $M = \lambda I - A$ by the spectral radius of the symmetric part of A if $A \geq 0$.

COROLLARY 5.2.2. *Let $M = \lambda I - A$ be a Z-matrix, where A is a nonnegative matrix and λ is a positive constant. If $\rho(H(A)) < \lambda$, then M is an M-matrix and*

$$\rho(M^{-1}) \leq \frac{1}{\lambda - \rho(H(A))}.$$

The next result estimates the bound of the numerical radius of matrices using the spectral radius of the Hermitian and anti-Hermitian parts.

COROLLARY 5.2.3. *Let A be an $n \times n$ matrix. Then*

$$\max(\rho(H(A)), \rho(S(H))) \leq r(A) \leq \sqrt{\rho^2(H(A)) + \rho^2(S(A))}.$$

Proof.

$$\mathbf{x}^* A\mathbf{x} = \mathbf{x}^* H(A)\mathbf{x} + \mathbf{x}^* S(A)\mathbf{x}.$$

It is easily seen that $\mathbf{x}^* H(A)\mathbf{x}$ is real and $\mathbf{x}^* S(A)\mathbf{x}$ is imaginary and the conclusion follows immediately. □

More generally, let P be a Hermitian positive definite matrix and let $(\mathbf{x}, \mathbf{y})_P$ denote an inner product on C^n defined by

$$(\mathbf{x}, \mathbf{y})_P = \mathbf{y}^* P\mathbf{x}.$$

Given a matrix A, the Rayleigh quotient of A for a vector $\mathbf{x} \neq 0$ w.r.t. the inner product $(\mathbf{x}, \mathbf{y})_P$ is

$$q_P(\mathbf{x}) = (A\mathbf{x}, \mathbf{x})_P / (\mathbf{x}, \mathbf{x})_P$$

and the numerical radius of A, denoted by $r_P(A)$, w.r.t. the inner product $(\mathbf{x}, \mathbf{y})_P$ is given by

$$r_P(A) = \sup\{|q_P(\mathbf{x})|; \mathbf{x} \in \mathbb{C}^n, \|\mathbf{x}\| \neq 0\}.$$

It is readily shown that $r_P(A)$ is a norm on the space of matrices. Let Q be a nonsingular matrix and define the matrix norm $\|\cdot\|_Q$ by

$$\|A\|_Q = \|QAQ^{-1}\|.$$

THEOREM 5.2.4. *Let $P = Q^*Q$ be Hermitian positive definite and $r_P(A)$ be the numerical radius w.r.t. the inner product $(\mathbf{x}, \mathbf{x})_P$, where Q is a nonsingular matrix. Then*

$$\begin{gathered} r_P(A) = r(QAQ^{-1}), \\ \rho(A) \leq r_P(A), \\ r_P(A^m) \leq r_P^m(A), \\ r_P(A) \leq \|A\|_Q \leq 2r_P(A). \end{gathered} \tag{5.2.4}$$

Proof. Since

$$\frac{(A\mathbf{x}, \mathbf{x})_P}{(\mathbf{x}, \mathbf{x})_P} = \frac{\mathbf{x}^* PA\mathbf{x}}{\mathbf{x}^* P\mathbf{x}} = \frac{\mathbf{y}^* QAQ^{-1}\mathbf{y}}{\mathbf{y}^*\mathbf{y}},$$

where $\mathbf{y} = Q\mathbf{x}$, (5.2.4) follows immediately. The remaining results follow from (5.2.4). □

From Theorem 5.2.1 and Theorem 5.2.4, the following result is immediate.

COROLLARY 5.2.5. *Let P be Hermitian positive definite, A be an $n \times n$ matrix and α be a constant such that $|\alpha| > r_P(A)$. Then $\alpha I - A$ is nonsingular and*

$$r_P((\alpha I - A)^{-1}) \leq \frac{1}{|\alpha| - r_P(A)}.$$

Proof. The proof is similar to that of Theorem 5.2.1. □

The following result shows that if we can choose a proper Hermitian positive definite matrix P, $r_P(A)$ estimates the spectral radius of A efficiently.

THEOREM 5.2.6. *For any square matrix A*

$$\inf_P r_P(A) = \rho(A),$$

where P is Hermitian positive definite.

Proof. The proof is essentially the same as that of Lemma 6.10 [9]. □

For other generalizations of the numerical radius see [13] and of fields of values, see [12]. The methods presented in [9] and [11] for determining fields of values can be useful for the numerical radius.

5.3. An Application to Basic Iterative Methods

Consider the basic

$$C\mathbf{x}^{(l+1)} = R\mathbf{x}^{(l)} + b, \qquad l = 0, 1, 2, \ldots, \tag{5.3.1}$$

for solving the linear system $A\mathbf{x} = \mathbf{b}$, where $A = C - R$ is a splitting of A and R is the defect matrix of the splitting. Let $B = C^{-1}R$ be the iteration matrix. Given a norm $\|\cdot\|$, $\|B^m\|$ is the convergence factor for m steps and $R_m = \|B^m\|^{\frac{1}{m}}$ is the average convergence factor for this norm. For any square matrix D with $\rho(D) > 0$, it is shown that there exist positive constants $c, C \in \mathbb{R}$ such that

$$cm^{s-1}\rho(D)^m \leq \|D^m\| \leq Cm^{s-1}\rho(D)^m, \quad m = 1, 2, \ldots, \tag{5.3.2}$$

where s is the order of the largest Jordan block that belongs to an eigenvalue λ with $|\lambda| = \rho(D)$ (see Young [19] (1971), Theorem 7.1, p85 for details).

It follows from (5.2.2) that

$$r(B^m)^{\frac{1}{m}} \leq \|B^m\|^{\frac{1}{m}} \leq 2^{\frac{1}{m}} r(B^m)^{\frac{1}{m}}. \tag{5.3.3}$$

(5.3.3) shows that $r(B^m)^{\frac{1}{m}}$ becomes an increasingly accurate estimate of $\|B^m\|^{\frac{1}{m}}$. Denote $\rho_M(A) = \max(\rho(H(A)), \rho(S(A)))$. (5.2.1), (3.3) and Proposition 5.2.3 show that

$$\begin{aligned} \rho_M(B^m)^{\frac{1}{m}} &\leq \|B^m\|^{\frac{1}{m}} \\ &\leq 2^{\frac{1}{m}}(r(B^m))^{\frac{1}{m}} \leq 2^{\frac{1}{m}} r(B) \leq 2^{\frac{1}{m}}\sqrt{2}\rho_M(B). \end{aligned} \tag{5.3.4}$$

In particular, if $B \geq 0$, (5.2.3) yields

$$\rho(H(B^m))^{\frac{1}{m}} \leq \|B^m\|^{\frac{1}{m}} \leq 2^{\frac{1}{m}}\rho(H(B)). \tag{5.3.5}$$

(5.3.2) implies that $\|B^m\|^{\frac{1}{m}} \to \rho(B)$. Hence, the average convergence factor (R_m) approaches asymptotically the spectral radius. Furthermore, $R_m \equiv \|B^m\|^{\frac{1}{m}} = \rho(B)$ for a symmetric matrix B. For unsymmetric matrices, however, $\rho(B)$ may be a by far too optimistic estimate of $\|B^m\|^{\frac{1}{m}}$ for practical values of m, because $\|B^m\|$ does not have to converge monotonically to zero and can even increase significantly for some initial values of m. Using the Jordan canonical form of B, $V^{-1}BV = J$, where V is a proper nonsingular transformation matrix and $J = \text{blockdiag}(J_1, J_2, \ldots, J_r)$, we can analyze the basic iteration $\mathbf{x}^{(l+1)} = B\mathbf{x}^{(l)} + \mathbf{b}$ as follows: Let $\mathbf{y}^{(l)} = V\mathbf{x}^{(l)}$ and $\tilde{\mathbf{b}} = V\mathbf{b}$. Then $\mathbf{x}^{(l+1)} = B\mathbf{x}^{(l)} + \mathbf{b}$ takes the form

$$\mathbf{y}^{(l+1)} = J\mathbf{y}^{(l)} + \tilde{\mathbf{b}}, \quad l = 0, 1, \ldots$$

Hence, to analyze the convergence behavior of $\mathbf{y}^{(l)}$ it suffices to consider $\|J_k^m\|^{\frac{1}{m}}$ for the Jordan blocks J_k, $k = 1, \ldots, r$. For simplicity we consider here only a Jordan block matrix of order 2. The main results about nonmonotone convergence will be similar for any order of the Jordan blocks.

Example 1. Nonmonotone convergence.

Consider the matrix $B = \begin{pmatrix} a & 1 \\ 0 & a \end{pmatrix}$, where $0 < a < 1$. We have $\rho(B) = a < 1$ which implies that the iterative method (5.3.1) converges if $B = C^{-1}R$. By induction we find

$$B^m = \begin{pmatrix} a^m & ma^{m-1} \\ 0 & a^m \end{pmatrix}, \; B^{mT}B^m = \begin{pmatrix} a^{2m} & ma^{2m-1} \\ ma^{2m-1} & a^{2m} + m^2a^{2m-2} \end{pmatrix}.$$

The largest eigenvalue λ_m of ${B^m}^T B^m$ is readily found to satisfy

$$\lambda_m \simeq a^{2m} + \frac{1}{2}m^2 a^{2m-2}\left(1 + \sqrt{1 + 4a^2/m^2}\right). \tag{5.3.6}$$

Since $\|B^m\| = \lambda_m^{\frac{1}{2}}$, (5.3.6) shows that $\|B^m\|$ increases until $m \simeq \frac{1}{4}\delta^{-1}$ for $a = 1-\delta$, $0 < \delta \ll 1$. There exists a certain positive integer m_0 such that $\|B^m\| \geq 1$ for all $m \leq m_0$. As an example, $m_0 = 644$ for $a = 0.99$.

On the other hand,

$$\begin{aligned} r(B^m) &= \rho(H(B^m)) = \rho(\frac{1}{2}(B^m + {B^m}^T)) \\ &= \rho\begin{pmatrix} a^m & ma^{m-1} \\ ma^{m-1} & a^m \end{pmatrix} = a^m(1 + \frac{m}{2a}) \end{aligned} \tag{5.3.7}$$

(5.3.5) and (5.3.7 show that

$$\|B^m\|^{\frac{1}{m}} \geq r(B^m)^{\frac{1}{m}} = a(1 + \frac{m}{2a})^{\frac{1}{m}} \geq 1$$

for $m \leq m_1$, where $m_1 = 562$, for $a = 0.99$.

In this case the spectral norm $\|J^m\|$ is computable, but the computational effort needed is larger than the effort to compute $r(J^m)$.

In (5.3.4) we have derived lower and upper bounds of the average convergence factor $\|B^m\|^{\frac{1}{m}}$ involving the numerical radius. The above examples show that the numerical radius is a more reliable function to use when analyzing iterative methods for nonsymmetric matrices than the spectral radius of the iteration matrix.

Let $e^{(m)} = \mathbf{x}^{(m)} - \mathbf{x}$. We have

$$\|e^{(m)}\| \leq \|B^m\|\|e^{(0)}\|.$$

Usually, $\|B^m\|$ is difficult to compute. Hence, in the case that $\|B\|$ is difficult to estimate or $\|B\| \geq 1$, it is hard to obtain a good rate of convergence. On the other hand, using the numerical radius, we have

$$\|e^{(m)}\| \leq \|B^m\|\|e^{(0)}\| \leq 2r(B^m)\|e^{(0)}\| \leq 2r^m(B)\|e^{(0)}\|.$$

Therefore, if $r(B) < 1$, which is certainly less restrictive than $\|B\| < 1$, we can easily derive an error estimate. This is illustrated in the next example.

Example 2. Consider the convection-diffusion equation,

$$\begin{aligned} &-\varepsilon u''(x) + bu'(x) - cu(x) = f(x), \quad 0 < x < 1, \\ &u(0) = \alpha, \quad u(1) = \beta, \end{aligned}$$

where ε, b and c are positive constants with $b \geq c$ and $\varepsilon > ch^2$. Discretizing the equation by an upwinded scheme with a uniform mesh $h = 1/(n+1)$, we obtain a linear system with a tridiagonal matrix of the form

$$A = \text{tridiag}(-\frac{\varepsilon}{h^2} - \frac{b}{h}, \frac{2\varepsilon}{h^2} + \frac{b}{h} - c, -\frac{\varepsilon}{h^2}).$$

Let $A = D - L - U$, where D is diagonal, L and U are strictly lower and upper triangular, respectively, and let $v = (1-(n-1)\delta, \ldots, 1-\delta, 1)^T$, where $\delta = ch/b$. One finds that $v > 0$ and $Av > 0$. Hence A is an M-matrix. Consider the forward

Gauss-Seidel method with iteration matrix $B_1 = (D - L)^{-1}U$. A computation shows that

$$\|B_1\|_\infty = \|B_1\|_1 = \left(1 - \left(\frac{\varepsilon + bh}{2\varepsilon + bh - ch^2}\right)^n\right)\frac{\varepsilon}{\varepsilon - ch^2},$$

$$r(B_1) \le \left(1 - \left(\frac{\varepsilon + bh}{2\varepsilon + bh - ch^2}\right)^{\frac{n}{2}}\right)\frac{\varepsilon}{\varepsilon - ch^2}.$$

Let $\gamma = \left(\frac{\varepsilon+bh}{2\varepsilon+bh-ch^2}\right)^n$. If $\gamma < \frac{ch^2}{\varepsilon} < \gamma^{\frac{1}{2}}$, then $\|B_1\|_\infty = \|B_1\|_1 > 1$. It is difficult to check if $\|B_1\| < 1$. In this case, however, $r(B_1) < 1$. We can use the numerical radius to estimate the rate of the convergence for any number of iteration steps. The spectral radius, which is

$$\rho(B_1) = \frac{4(bh + \varepsilon)\varepsilon}{(2\varepsilon + bh - ch^2)^2}\cos^2\pi h,$$

gives only the asymptotic rate of convergence.

Examples of using fields of values to estimate the convergence of iterative methods can be found in [6] and [15].

5.4. A Use in the Analysis of the SOR Method

Let $A = D - U - L$, where $D = \mathrm{diag}(D_1, D_2, \ldots, D_r)$ is the block diagonal part of A. L and U are the lower and upper block triangular parts of A respectively. Consider the SOR method

$$(\frac{1}{\omega}D - L)\mathbf{x}^{(l+1)} = [(\frac{1}{\omega} - 1)D + U]\mathbf{x}^{(l)} + b, \qquad l = 0, 1, \ldots,$$

where $\omega \neq 0$ is a relaxation parameter. The iteration matrix is

$$L_\omega = (\frac{1}{\omega}D - L)^{-1}((\frac{1}{\omega} - 1)D + U). \tag{5.4.1}$$

The matrix A is said to be consistently ordered if $\alpha D^{-1}L + \alpha^{-1}D^{-1}U$, $\alpha \neq 0$, has eigenvalues which do not depend on α. It is well known that there exists a unique value ω_{opt} for consistently ordered matrices such that $\rho(L_{\omega_{opt}}) < \rho(L_\omega)$, $\forall\omega \neq \omega_{opt}$ (see [16], [17], [18], [19] for details). As an application of the numerical radius, we now investigate some properties of the SOR method for so called quasi-Hermitian positive definite matrices without requiring that the matrices are consistently ordered.

THEOREM 5.4.1. *Let $A = D - L - U$ and L_ω be the iteration matrix defined by* (5.4.1) *with $0 < \omega < 2$. If there exists a nonsingular block diagonal matrix $P = \mathrm{diag}(P_1, P_2, \ldots, P_r)$ such that $\tilde{A} = PAP^{-1}$ is Hermitian positive definite, i.e., A is a quasi-Hermitian matrix, where the order of P_i is equal to that of D_i, then*

$$\rho(L_\omega)^2 \le 1 - \frac{\frac{2}{\omega} - 1}{(\frac{1}{\omega} - \frac{1}{2})^2\delta^{-1} + \gamma + \frac{1}{\omega}}, \tag{5.4.2}$$

where

$$\gamma = \sup_{\mathbf{x}\neq 0}\left\{\left[\frac{|(\mathbf{x}, \tilde{L}\mathbf{x}|^2}{(\mathbf{x},\mathbf{x})} - \frac{1}{4}(\mathbf{x},\mathbf{x})\right]/(\tilde{A}\mathbf{x},\mathbf{x})\right\} \le (r(\tilde{L})^2 - \frac{1}{4})/\delta,$$

$$\delta = \lambda_{\min}(\tilde{A}) = \min_{\mathbf{x}\neq 0}(\tilde{A}\mathbf{x},\mathbf{x})/(\mathbf{x},\mathbf{x})$$

and

$$G^*G = PDP^{-1},\ \tilde{A} = G^{*^{-1}}\bar{A}G^{-1},\ \tilde{L} = G^{*^{-1}}PLP^{-1}G^{-1}.$$

Furthermore, if $r(\tilde{L}) \leq \frac{1}{2}$, *then* $\gamma \leq 0$ *and* $\omega^* = 2/(1+\sqrt{2\delta})$ *minimizes the upper bound in* (5.4.2) *of* $\rho(L_\omega)$ *and we have*

$$\rho(L_{\omega^*})^2 = (1-\sqrt{\delta/2})/(1+\sqrt{\delta/2})$$

Proof. Note that under the assumptions of the theorem, PDP^{-1} is Hermitian positive definite and $\bar{U} = \bar{L}^*$, where $\bar{L} = PLP^{-1}$ and $\bar{U} = PUP^{-1}$. Transforming L_ω by GP, we have

$$\begin{aligned}
\tilde{L}_\omega &= GP(\frac{1}{\omega}D - L)^{-1}P^{-1}G^*G^{*^{-1}}P((\frac{1}{\omega}-1)D+U)P^{-1}G^{-1} \\
&= G(\frac{1}{\omega}\bar{D} - \bar{L})^{-1}G^*G^{*^{-1}}((\frac{1}{\omega}-1)\bar{D}+\bar{U})G^{-1} \\
&= (\frac{1}{\omega}I - \tilde{L})^{-1}((\frac{1}{\omega}-1)I + \tilde{L}^*)
\end{aligned}$$

Let $(\lambda, \mathbf{x})$ be an eigenpair of $\tilde{L}_\omega$ and $z = (\mathbf{x}, \tilde{L}\mathbf{x})/(\mathbf{x},\mathbf{x})$. One finds

$$(\frac{1}{\omega}-1)I + \tilde{L}^*)\mathbf{x} = \lambda(\frac{1}{\omega}I - \tilde{L})\mathbf{x}.$$

A simple computation shows

$$\lambda = (\frac{1}{\omega} - 1 + \bar{z})/(\frac{1}{\omega} - z).$$

Hence

$$\begin{aligned}
|\lambda|^2 &= \frac{\left((\frac{1}{\omega}-1)^2 + 2(\frac{1}{\omega}-1)Re(z) + |z|^2 \right)}{\left((\frac{1}{\omega})^2 - \frac{2}{\omega}Re(z) + |z|^2 \right)} \\
&= 1 - \frac{(\frac{2}{\omega}-1)(1-2Re(z))}{\left((\frac{1}{\omega}-\frac{1}{2})^2 + \frac{1}{\omega}(1-2Re(z)) + |z|^2 - \frac{1}{4} \right)}.
\end{aligned}$$

Since $\tilde{A} = I - \tilde{L} - \tilde{L}^*$, we have $(\tilde{A}\mathbf{x},\mathbf{x})/(\mathbf{x},\mathbf{x}) = 1 - (z+\bar{z}) = 1 - 2Re(z)$. Thus

$$|\lambda|^2 = 1 - \ (\tfrac{2}{\omega}-1)\ \Big/ \left\{ \left((\tfrac{1}{\omega}-\tfrac{1}{2})^2 + \left| \tfrac{(\mathbf{x},\tilde{L}\mathbf{x})}{(\mathbf{x},\mathbf{x})} \right|^2 - \tfrac{1}{4} \right) \tfrac{(\mathbf{x},\mathbf{x})}{(\tilde{A}\mathbf{x},\mathbf{x})} + \tfrac{1}{\omega} \right\},$$

which shows (5.4.2). If $r(\tilde{L}) \leq \frac{1}{2}$, then $\gamma \leq 0$ and

$$\rho(\tilde{L}_\omega)^2 \leq 1 - 2\Big/ \left((1-\tfrac{\omega}{2})\tfrac{1}{\omega\delta} + (1-\tfrac{\omega}{2})^{-1} \right) .$$

It is readily seen that $\omega = \omega^* = 2/(1+\sqrt{2\delta})$ minimizes this upper bound and for this value of ω we get $\rho(\tilde{L}_{\omega^*})^2 \leq (1-\sqrt{\delta/2})/(1+\sqrt{\delta/2})$. □

Regarding the constant γ, here we have the following result.

PROPOSITION 5.4.2. *Let the conditions of* Theorem 5.4.1 *hold. Then*

$$\gamma = \frac{1}{4}\sup_{\mathbf{x}\neq 0}\frac{|(\mathbf{x},(\tilde{L}-\tilde{L}^*)\mathbf{x})|^2}{(\mathbf{x},\tilde{A}\mathbf{x})}.$$

Proof. Let $z = (\mathbf{x}, \tilde{L}\mathbf{x})$. As proved in Proposition 5.2.3,

$$|z|^2 = \frac{1}{4}(\mathbf{x},(\tilde{L}+\tilde{L}^*)\mathbf{x})^2 + \frac{1}{4}(\mathbf{x},(\tilde{L}-\tilde{L}^*)\mathbf{x})^2.$$

Because $\tilde{A} = I - \tilde{L} - \tilde{L}^*$ is positive definite, $(\mathbf{x}, (\tilde{L} + \tilde{L}^*)\mathbf{x}) \leq 1$ and

$$\gamma = \sup_{(\mathbf{x},\mathbf{x})=1} (|z|^2 - \frac{1}{4})/(\tilde{A}\mathbf{x}, \mathbf{x}) \leq \sup_{(\mathbf{x},\mathbf{x})=1} (\mathbf{x}, (\tilde{L} - \tilde{L}^*)\mathbf{x})^2/(\tilde{A}\mathbf{x}, \mathbf{x}),$$

which finishes the proof. □

REMARK 5.4.3. Theorem 5.4.1 is applicable to the case that A is not consistently ordered and not a Stieltjes matrix. If A is consistently ordered

$$\rho(L_{\omega_{opt}}) = \omega_{opt} - 1,$$

and if A is a Stieltjes matrix,

$$\omega_{opt} - 1 \leq \rho(L_{\omega_{opt}}) \leq (\omega_{opt} - 1)^{\frac{1}{2}},$$

see [17], [19].

REMARK 5 4.4. If $\tilde{L} \geq 0$, (5.2.3) shows that

$$r(\tilde{L}) = \frac{1}{2} \max_{\substack{\mathbf{x} \in \mathbf{R}^n \\ (\mathbf{x},\mathbf{x}) = 1}} \mathbf{x}^T(\tilde{L} + \tilde{L}^T)\mathbf{x},$$

so $r(\tilde{L}) \leq \frac{1}{2}$ and by Theorem 5.4.1, $\gamma \leq 0$.

REMARK 5.4.5. A formula similar to (5.4.2) was first derived in [1] and [4]. It is interesting to note that in a similar expression for the SSOR method, the corresponding constant γ involves the spectral radius of $\tilde{L}\tilde{L}^T$, see [2] and [19].

5.5. Upper Eigenvalue Bounds of ILU Preconditioners

Let A be a symmetric matrix split as

$$A = D_A + L_A + L_A^T, \tag{5.5.1}$$

where D_A and L_A are the block diagonal part and the lower block triangular part of A, respectively. Consider a block incomplete preconditioner of the form

$$C = (X + L)X^{-1}(X + L), \tag{5.5.2}$$

where X is block diagonal and s.p.d. matrix and L is a block lower triangular matrix. In addition, X and L are partitioned in blocks consistently with D_A and L_A respectively.

To estimate the rate of convergence of preconditioned iterative methods such as the Chebyshev iterative method and (generalized) conjugate gradient method we need to know the extreme eigenvalues of the preconditioned matrix. As the third application of numerical radius in this paper, we use it to estimate the upper bound of eigenvalues of $C^{-1}A$. For later use of this section, we state the following result given by Axelsson and Lu [5] without proof.

THEOREM 5.5.1. *Let A be a symmetric matrix and $C = (X+L)X^{-1}(X+L^T)$ defined by* (5.5.1) *and* (5.5.2) *respectively, where X is s.p.d. If there exist two constants σ and β such that $\beta X \leq K \leq \sigma X$ in a positive semidefinite sense, then*

$$\lambda_i(M(\beta)) \leq \lambda_i(C^{-1}A) \leq \lambda_i(M(\sigma)),$$

where $\lambda_i(D) \leq \lambda_{i+1}(D)$, $i = 1, 2, \ldots, n-1$, denote the eigenvalues of D, $K = A - L - L^T$, and

$$M(s) = (I + \tilde{L})^{-1} + (I + \tilde{L}^T)^{-1} + (s-2)(I + \tilde{L})^{-1}(I + \tilde{L}^T)^{-1},$$
$$\tilde{L} = X^{-\frac{1}{2}} L X^{-\frac{1}{2}}.$$

Based on this result, we can estimate the maximum eigenvalue of $C^{-1}A$ by the numerical radius. The aim here is to investigate how $\tilde{L}$ influences the upper bound of eigenvalues of $C^{-1}A$.

THEOREM 5.5.2. *Let A be partitioned in $m \times m$ blocks and assume that the conditions of* Theorem 5.5.1 *hold. If $\sigma \leq 2$ and $r(\tilde{L}) \leq 1 + \frac{c}{m}$ for some nonnegative constant c, which does not depend on m, then*

$$\lambda_{\max}(C^{-1}A) \leq \begin{cases} 2m\frac{e^c-1}{c}, & \text{if } c > 0 \\ 2m, & \text{if } c = 0 \end{cases}.$$

Proof. Applying Theorem 5.5.1, Proposition 5.2.3 and (5.2.1) shows that

$$\begin{aligned}
\lambda_{\max}(C^{-1}A) &\leq \rho((I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1}) \\
&= \max_{\mathbf{x}^T\mathbf{x}=1} |\mathbf{x}^T((I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1})\mathbf{x}| \\
&= \max_{\mathbf{x}^T\mathbf{x}=1} \left| \mathbf{x}^T \left(\sum_{k=0}^{m-1} (-\tilde{L})^k + \sum_{k=0}^{m-1} (-\tilde{L})^{k^T} \right) \mathbf{x} \right| \\
&\leq 2 \max_{\mathbf{x}^T\mathbf{x}=1} \sum_{k=0}^{m-1} \left| \mathbf{x}^T \frac{\tilde{L}^k + \tilde{L}^{k^T}}{2} \mathbf{x} \right| \leq 2 \sum_{k=0}^{m-1} \rho(H(\tilde{L}^k)) \\
&\leq 2 \sum_{k=0}^{m-1} r(\tilde{L}^k) \leq 2 \sum_{k=0}^{m-1} r^k(\tilde{L}) \leq \begin{cases} 2m\frac{e^c-1}{c}, & \text{if } c > 0, \\ 2m, & \text{if } c = 0, \end{cases}
\end{aligned}$$

which is the result desired. □

Note that, firstly, the power inequality (5.2.1) is being used here, secondly, that if one used the spectral radius one would have to bound $\rho(\tilde{L}^k + \tilde{L}^{k^T})$ which may not be easy, and finally that if one used the spectral norm then the bound could be considerably worse.

Theorem 5.5.2 shows, in particular, that if $r(\tilde{L}) \leq 1$, then $\lambda(C^{-1}A) \leq 2m$. Our next result estimates the upper bound of eigenvalues in the case of $r(\tilde{L}) < 1$.

THEOREM 5.5.3. *Let A be partitioned in $m \times m$ blocks and assume that the conditions of* Theorem 5.5.1 *hold. If $r(\tilde{L}) \leq 1 - \varepsilon$, where $0 < \varepsilon < 1$, then*

$$\lambda_{\max}(C^{-1}A) \leq \begin{cases} 2\varepsilon^{-1}, & \text{if } \sigma \leq 2, \\ 2\varepsilon^{-1} + (\sigma-2)\varepsilon^{-2}, & \text{if } \sigma > 2. \end{cases}$$

Proof. Note that $r(B) = r(B^*)$ for any square matrix B. If $\sigma \leq 2$, Theorem 5.5.1 and Theorem 5.2.1 show

$$\begin{aligned}
\lambda_{\max}(C^{-1}A) &\leq \lambda_{\max}((I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1}) \\
&= \max_{\mathbf{x}^T\mathbf{x}=1} \mathbf{x}^T((I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1})\mathbf{x} \leq r((I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1}) \\
&\leq r((I+\tilde{L})^{-1}) + r((I+\tilde{L}^T)^{-1}) \leq \frac{1}{1-r(\tilde{L})} + \frac{1}{1-r(\tilde{L}^T)} \leq 2\varepsilon^{-1}
\end{aligned}$$

Note that for any two symmetric matrices B and D

$$\lambda_{\max}(B+D) \leq \lambda_{\max}(B) + \lambda_{\max}(D).$$

Hence, If $\sigma > 2$, Theorem 5.5.1 shows

$$\begin{aligned}
\lambda_{\max}(C^{-1}A) \quad \leq \quad & \lambda_{\max}((I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1}) \\
& + (\sigma-2)\lambda_{\max}((I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1}).
\end{aligned}$$

Therefore, it remains only to prove

$$\lambda_{\max}((I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1}) \leq \varepsilon^{-2}.$$

To this end, Let $(\lambda, \mathbf{x})$ be an eigenpair of $(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1}$. We have

$$(I+\tilde{L}^T)^{-1}\mathbf{x} = \lambda(I+\tilde{L})\mathbf{x},$$

which implies

$$|\lambda|(1-r(\tilde{L})) \leq |\lambda|\left|\frac{\mathbf{x}^T(I+\tilde{L})\mathbf{x}}{\mathbf{x}^T\mathbf{x}}\right| = \left|\frac{\mathbf{x}^T(I+\tilde{L}^T)^{-1}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}\right| \leq r((I+\tilde{L}^T)^{-1}).$$

Theorem 5.2.1 now shows

$$|\lambda| \leq \frac{r((I+\tilde{L}^T)^{-1})}{1-r(\tilde{L})} \leq \frac{1}{(1-r(\tilde{L}))^2} \leq \varepsilon^{-2},$$

which ends the proof. □

If $\sigma < 2$ it is shown in [5] that $\lambda_{\max}(C^{-1}A) \leq \frac{1}{2-\sigma}$ for symmetric matrix A, but $\frac{1}{2-\sigma}$ is still very large if σ is close to 2. If A is positive definite, X is a Stieltjes matrix and L is nonpositive, the following improved result holds.

THEOREM 5.5.4. *Let A be a symmetric positive definite matrix partitioned in $m \times m$ blocks and $C = (X+L)X^{-1}(X+L^T)$ defined by* (5.5.1) *and* (5.5.2) *respectively, where X is a Stieltjes matrix and L is nonpositive. If there exists a constant $\sigma < 2$ such that $\sigma X - K$ is positive semidefinite, where $K = A - L - L^T$, then*

$$\lambda_{\max}(C^{-1}A) \leq \min\left\{\frac{2}{2-\sigma+\lambda_{\min}(\tilde{A})}, 2m\right\},$$

where $\tilde{A} = X^{-\frac{1}{2}}AX^{-\frac{1}{2}}$.

Proof. Since X is a Stieltjes matrix, $X^{-\frac{1}{2}}$ is nonnegative so $\tilde{L} = X^{-\frac{1}{2}}LX^{-\frac{1}{2}}$ is nonpositive. Under the assumption of the theorem,

$$\sigma X + L + L^T = (\sigma X - K) + A.$$

Hence,

$$\sigma I + \tilde{L} + \tilde{L}^T = X^{-\frac{1}{2}}(\sigma X - K)X^{-\frac{1}{2}} + \tilde{A},$$

which, using the assumption that $\sigma X - K$ is positive semidefinite, implies

$$\max_{\mathbf{x}^T\mathbf{x}=1} \mathbf{x}^T\frac{1}{2}(-\tilde{L}-\tilde{L}^T)\mathbf{x} \leq \frac{1}{2}(\sigma - \lambda_{\min}(\tilde{A})).$$

(5.2.3) shows now

$$r(\tilde{L}) \leq \frac{1}{2}(\sigma - \lambda_{\min}(\tilde{A})) = 1 - \frac{1}{2}(2-\sigma+\lambda_{\min}(\tilde{A})).$$

Theorem 5.5.2 and Theorem 5.5.3 end the proof. □

Acknowledgments

Some useful comments about the paper by an anonymous referee are gratefully acknowledged.

References

[1] L. ANDERSON, *SSOR preconditioning of Toeplitz matrices*, Ph.d thesis, Department of Computer Science, Chalmers University of Technology, Göteborg, Sweden, 1976.

[2] O. AXELSSON, *A generalized SSOR methods*, BIT, 13 (1972), pp. 443–467.

[3] ———, *On preconditioning and convergence acceleration in sparse matrix problems*, CERN Technical Report, 74-10 (1974), Data Handling Division, Geneva.

[4] ———, *Solution of linear systems of equations: iterative methods*, in Sparse Matrix Techniques, Lecture Notes in Mathematics 572, V. A. Barker, ed., Springer Verlag, Berlin, Heidelberg, New York, 1977, pp. 1–50

[5] O. AXELSSON AND H. LU, *On eigenvalue estimates for block incomplete factorization methods*, SIAM J. Matrix Anal. Appl., (to appear).

[6] M. EIERMANN, *Fields of values and iterative methods*, Linear Algebra Appl., 180 (1993), pp. 167–197

[7] M. GOLDBERG AND E TADMOR, *On the numerical radius and its applications*, Lin. Alg. Appl., 42 (1982), pp. 263–284.

[8] M. GOLDBERG, E. TADMOR AND G. ZWAS, *Numerical radius of positive matrices*, Lin. Alg. Appl., 12 (1975), pp. 209–214.

[9] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[10] ———, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[11] C. R. JOHNSON, *Numerical determination of the field of values of a general complex matrix*, SIAM J. Numer. Anal. 15 (1978), pp. 595–602.

[12] H. W. J. LENFERINK AND M. N. SPIJKER, *A generalization of the numerical range of a matrix*, Linear Algebra Appl. 140 (1990), pp. 251–266.

[13] C. K. LI, *A generalization of spectral radius, numerical radius and spectral norm*, Linear Algebra Appl , 90 (1987), pp. 105–118.

[14] C. PEARCY, *An elementary proof of the power inequality for the numerical radius*, Michigan Math. J 13 (1966), pp 289–291.

[15] G. STARKE, *Fields of values and the ADI method for nonnormal matrices*, Linear Algebra Appl. 180 (1993), pp. 199–218

[16] R. VARGA, *Ordering of successive overrelaxation schemes*, Pacific J. Math., 9 (1959), pp. 925 936.

[17] ———, *Matrix Iterative Analysis*, Prentice Hall, Englewood Cliffs, N.J., 1962.

[18] D. M. YOUNG, *Iterative methods for solving partial difference equations of elliptic type*, Ph.d thesis, Harvard University, 1950.

[19] ———, *Iterative Solution of Large Systems*, Academic Press, New York, 1971.

CHAPTER 6

Matrix Compensation and Diagonal Compensation*

Abstract. Matrix compensation is introduced to preserve the positivity of matrices when we construct preconditioners for positive definite matrices. Diagonal compensation for symmetric positive definite matrices is generalized to positive definite matrices. If matrix compensation is used to construct preconditioners, condition numbers of preconditioned matrices are estimated.

Key words. matrix compensation, diagonal compensation, eigenvalue estimates, preconditioners

AMS subject classifications. 65F10, 65F15, 65F50

6.1. Introduction

When we construct a preconditioner for a positive definite matrix, it is important to preserve the positivity of matrices. Let A be a positive definite matrix. Roughly speaking, matrix compensation is to find a matrix G such that $G - A$ is positive semidefinite. In general, it is not easy to find a good compensative matrix G to construct a good preconditioner for A by using G. In some occasions, however, the diagonal compensation can be used to obtain matrix compensation for symmetric positive definite matrices easily for constructing good preconditioners [2], [3], [4]. The idea of diagonal compensation can be traced back to Axelsson's early work [1]. Further study can be found in [2], [3], [4], [5], [9]. It is well known that diagonal compensation acts as a key in modified block incomplete factorization. We generalize the diagonal compensation to positive definite matrices.

Let $A = B + R$ be a symmetric positive definite matrix and $M = B + D$, where B is a symmetric matrix and D is a diagonally compensative matrix of R. M can be an efficient preconditioner of A (see e.g. [2], [3], [4]). If $Bv > 0$ for some positive vector v or $\rho(A^{-1}R) < 1$, some general results on upper bounds of eigenvalues and condition number for $M^{-1}A$ are derived by using the spectral radius $\rho(B^{-1}R)$ and the condition number $\kappa(M^{-1}B)$ in [2], [4]. Sometimes, however, $\rho(B^{-1}R)$ and $\kappa(M^{-1}B)$ are difficult to estimate accurately, where and throughout the paper $\kappa(C^{-1}A) = \lambda_{\max}(C^{-1}A)/\lambda_{\min}(C^{-1}A)$ for symmetric positive definite matrices A and C. We show here some new upper bounds of eigenvalues and condition numbers for compensative preconditioners involving $\rho(B^{-1}D)$ or $\rho(M^{-1}D)$. Finally, condition numbers of preconditioned matrices are estimated if matrix compensation is used to construct preconditioners for symmetric positive definite matrices.

---

* This chapter is based on the paper, H. Lu, *Matrix compensation and diagonal compensation*, J. Comp. Appl. Math., (to appear).

6.2. Matrix Compensation and Diagonal Compensation

In this section we define matrix compensation and generalize diagonal compensation to positive definite matrices.

DEFINITION A matrix D is called a compensative matrix of R if $D - R$ is positive semidefinite.

To make matrix compensation simple for a given positive definite matrix A, we choose a matrix R for which one can easily find a compensative matrix and consider the matrices

$$A = B + R, \tag{6.2.1}$$

$$M = B + D, \tag{6.2.2}$$

where D is a compensative matrix of R. Clearly M is a positive definite matrix and a compensative matrix of A. For practical reason, frequently B and D are assumed to be sparse matrices or matrices with special structures.

Let $A = (a_{i,j})_{i,j=1}^n$ and S be a sparse pattern. We can find a compensative matrix M with the pattern S as follows. Let $(B)_{i,j} = a_{i,j}$ if $(i,j) \in S$, $(B)_{i,j} = 0$ if $(i,j) \notin S$, and $R = A - B$. Choosing a compensative matrix D with the pattern S of matrix R, one finds that $M = B + D$ is a result.

We now generalize the diagonal compensation to positive definite matrices. Let A be a positive definite matrix split as (6.2.1) and denote $|R| = (|r_{i,j}|)_{i,j=1}^n$ for $R = (r_{i,j})_{i,j=1}^n$. Assume D is a diagonal matrix such that

$$D\mathbf{v} \geq \left|\frac{R + R^T}{2}\right|\mathbf{v} \tag{6.2.3}$$

for a positive vector $\mathbf{v} = (v_1, v_2, \ldots, v_n)^T$. It is not hard to see from (6.2.3) that matrix $V^{-1}(D - \frac{R+R^T}{2})V$ is diagonally dominant, where $V = \mathrm{diag}(v_1, v_2, \ldots, v_n)$. The Gerschgorin Circle Theorem (see e.g. [8]) shows that $\lambda(V^{-1}(D - \frac{R+R^T}{2})V)$ is nonnegative, so is $\lambda(D - \frac{R+R^T}{2})$. Hence, $D - \frac{R+R^T}{2}$ is symmetric positive semidefinite. On the other hand,

$$\mathbf{x}^T(D - R)\mathbf{x} = \mathbf{x}^T\left(D - \frac{R + R^T}{2}\right)\mathbf{x}, \quad \forall \mathbf{x} \in \mathbb{R}^n,$$

which implies that $D - R$ is positive definite, i.e., D is a compensative matrix of R. We call D a diagonally compensative matrix of R.

If A is a symmetric positive definite matrix and B is a symmetric matrix, (6 2.3) becomes $D\mathbf{v} \geq |R|\mathbf{v}$ so it is seen that (6.2.3) generalizes diagonal compensation for symmetric positive definite matrices [2], [3], [4]. In this case M given by (6.2.2) is also symmetric positive definite and $\lambda(M^{-1}A) \leq 1$.

Elsner and Mehrmann presented a class of $m \times m$ block matrices $A = (A_{i,j})_{i,j=1}^m$ [7], i.e., a generalization of Z-matrices, where the blocks $A_{i,j}$'s are $k \times k$ Hermitian matrices and the off-diagonal blocks are negative semidefinite (see also Axelsson [2]). These matrices arise for example in the numerical solution of Euler equations [7], [10]. Denote $S_m^k = \{A = (A_{i,j})_{i,j=1}^m \mid A_{i,j}$'s are $k \times k$ symmetric matrices $\}$. If $A = (A_{i,j})_{i,j=1}^m \in S_m^k$, split $A = B + R$, where $B, R = (R_{i,j})_{i,j=1}^m \in S_m^k$. and $R_{ii} = 0$, $i = 1, 2, \ldots, m$. Let D=blockdiag$(D_1, D_2, \ldots, D_n)$ be a block diagonal matrix such

that

$$2D_iu_i - \sum_{j=1}^{m} u_j(|R_{ij}|_2 + |R_{ji}|_2), \quad i = 1, 2, \ldots, m$$

are positive semidefinite for some positive vector $\mathbf{u} = (u_1, u_2, \ldots, u_m)^T \in \mathbb{R}^m$, where $|G|_2 = (GG)^{\frac{1}{2}}$ and D_i's are $k \times k$ symmetric positive definite matrices. It follows from [10] that $D - R$ is positive semidefinite. Hence, D is a compensative matrix of R. We call D a block diagonally compensative matrix of R. $M = B + D$ becomes clearly a compensative matrix of A. In the rest of the paper, we assume that all matrices are symmetric.

6.3. Analysis of Compensative Preconditioners

Given a matrix A, a compensative preconditioner means using a compensative matrix of A as a preconditioner of A. In this section, we will analyze this kind of preconditioners. Some analysis and practical examples of compensative preconditioners can be found in [2], [4].

THEOREM 6.3.1. *Let $A = B + R$ and $M = B + D$, where D is a compensative matrix of R. Assume that A and B are positive definite. Then*

(a) *If D is positive semidefinite, $\kappa(M^{-1}B) \leq (1-\rho(M^{-1}D))^{-1} = 1+\rho(B^{-1}D)$.*

(b) *If D is a diagonally compensative matrix of R and $\rho(B^{-1}D) < 1$,*

$$\kappa(M^{-1}A) \leq \frac{1+\rho(B^{-1}D)}{1-\rho(B^{-1}D)} = \frac{1}{1-2\rho(M^{-1}D)}.$$

Proof. To prove the theorem we need the following inequalities (see [2], [4])

$$\lambda_i(\widetilde{M}^{-1}\widetilde{B})\lambda_{\max}(\widetilde{B}^{-1}\widetilde{A}) \geq \lambda_i(\widetilde{M}^{-1}\widetilde{A}) \geq \lambda_i(\widetilde{M}^{-1}\widetilde{B})\lambda_{\min}(\widetilde{B}^{-1}\widetilde{A}), \tag{6.3.1}$$

where $\widetilde{A}$, $\widetilde{B}$ and $\widetilde{M}$ are positive definite.

If D is positive semidefinite, $M = B + D$ implies that $\rho(M^{-1}B) \leq 1$ and $\rho(M^{-1}D) < 1$. Hence,

$$\begin{aligned} \kappa(M^{-1}B) &\leq (\lambda_{\min}(M^{-1}B))^{-1} \\ &= (\lambda_{\min}(I - M^{-1}D))^{-1} = (1 - \rho(M^{-1}D))^{-1}. \end{aligned}$$

On the other hand, $\lambda(B^{-1}M) = \lambda^{-1}(M^{-1}B) \geq 1$ shows that

$$\begin{aligned} 1 + \rho(B^{-1}D) &= 1 + \rho(B^{-1}(M - B)) = \rho(B^{-1}M) \\ &= (\lambda_{\min}(M^{-1}B))^{-1} = (1 - \rho(M^{-1}D))^{-1}, \end{aligned}$$

which ends the proof of (a).

If D is a diagonally compensative matrix of R, i.e., D is a nonnegative diagonal matrix such that $D\mathbf{v} \geq \left|\frac{R+R^T}{2}\right|\mathbf{v} = |R|\mathbf{v}$ for some positive vector $\mathbf{v}$, then as mentioned above, both $D - R$ and $D + R$ are positive semidefinite, so are $B^{-\frac{1}{2}}DB^{-\frac{1}{2}} - B^{-\frac{1}{2}}RB^{-\frac{1}{2}}$ and $B^{-\frac{1}{2}}DB^{-\frac{1}{2}} + B^{-\frac{1}{2}}RB^{-\frac{1}{2}}$. Hence, $\rho(B^{-1}D) \geq \rho(B^{-1}R)$. $0 < \lambda(M^{-1}A) \leq 1$, $\rho(B^{-1}D) < 1$ and (6.3.1) show that

$$\begin{aligned} \kappa(M^{-1}A) &\leq (\lambda_{\min}(M^{-1}A))^{-1} \leq (\lambda_{\min}(M^{-1}B)\lambda_{\min}(B^{-1}A))^{-1} \\ &= \frac{1+\rho(B^{-1}D)}{1+\lambda_{\min}(B^{-1}R)} \leq \frac{1+\rho(B^{-1}D)}{1-\rho(B^{-1}R)} \leq \frac{1+\rho(B^{-1}D)}{1-\rho(B^{-1}D)}. \end{aligned}$$

The last equality of (b) follows from the equality $(1-\rho(M^{-1}D))^{-1} = 1+\rho(B^{-1}D)$. □

Other estimates of upper bounds for eigenvalues and condition numbers of $M^{-1}A$ by using $\rho(B^{-1}R)$ and $\kappa(M^{-1}B)$ can be found in [2], [4].

The theorem requires that B is positive definite. The following result shows a simple condition for both $\rho(B^{-1}D) < 1$ and B being positive definite. In particular, $\rho(B^{-1}D) < 1$ implies that B is positive definite if D is positive semidefinite.

LEMMA 6.3.2. *Let $M = B + D$ be positive definite. If $\rho(M^{-1}D) < \frac{1}{2}$, then B is positive definite and*

$$\rho(B^{-1}D) < 2\rho(M^{-1}D). \tag{6.3.2}$$

If, in addition, D is a positive semidefinite matrix, then $\rho(B^{-1}D) < 1$ if and only if $\rho(M^{-1}D) < \frac{1}{2}$.

Proof. $M = B + D$ and $\rho(M^{-1}D) < \frac{1}{2}$ show that $\lambda(M^{-1}B) \geq 1 - \rho(M^{-1}D) > \frac{1}{2}$, which implies that B is positive definite and $0 < \lambda(B^{-1}M) < 2$. A simple computation shows that $-1 < \lambda(B^{-1}D) = \lambda(B^{-1}M) - 1 < 1$, i.e., $\rho(B^{-1}D) < 1$. Thus, $I + B^{-\frac{1}{2}}DB^{-\frac{1}{2}}$ is positive semidefinite and

$$\begin{aligned}
&\rho(M^{-1}D) = \rho(B^{-\frac{1}{2}}(I + B^{-\frac{1}{2}}DB^{-\frac{1}{2}})^{-1}B^{-\frac{1}{2}}D) \\
= \;& \rho((I + B^{-\frac{1}{2}}DB^{-\frac{1}{2}})^{-\frac{1}{2}}B^{-\frac{1}{2}}DB^{-\frac{1}{2}}(I + B^{-\frac{1}{2}}DB^{-\frac{1}{2}})^{-\frac{1}{2}}) \\
= \;& \max_{\mathbf{x}\neq 0}\left|\frac{\mathbf{x}^T B^{-\frac{1}{2}}DB^{-\frac{1}{2}}\mathbf{x}}{\mathbf{x}^T(I + B^{-\frac{1}{2}}DB^{-\frac{1}{2}})\mathbf{x}}\right| \geq \max_{\mathbf{x}^T\mathbf{x}=1}\frac{|\mathbf{x}^T B^{-\frac{1}{2}}DB^{-\frac{1}{2}}\mathbf{x}|}{1 + |\mathbf{x}^T B^{-\frac{1}{2}}DB^{-\frac{1}{2}}\mathbf{x}|} \\
\geq \;& \frac{\rho(B^{-1}D)}{1 + \rho(B^{-1}D)} > \frac{1}{2}\rho(B^{-1}D),
\end{aligned}$$

which completes (6.3.2).

If D is positive semidefinite, we have

$$0 \leq \lambda(M^{-1}D) = 1 - \lambda(M^{-1}B) = 1 - \lambda^{-1}(B^{-1}M). \tag{6.3.3}$$

On the other hand $\rho(B^{-1}D) < 1$ shows $\lambda(B^{-1}M) = \lambda(I + B^{-1}D) \leq 1 + \rho(B^{-1}D) < 2$. Equation (6.3.3) shows that $\rho(M^{-1}D) < \frac{1}{2}$. □

Because M, B and D are often sparse matrices, we estimate the upper bounds of $\lambda_i(M^{-1}A)$ by using $\lambda_i(M^{-1}B)$ and $\rho(M^{-1}D)$.

THEOREM 6.3.3. *Let $A = B + R$ and $M = B + D$, where D is a compensation matrix of R. If B is positive definite and A and D are positive semidefinite, then*

$$\begin{aligned}
&(1 - \rho(M^{-1}D))^{-1}\lambda_i(M^{-1}B) \\
&\qquad = (1 + \rho(B^{-1}D))\lambda_i(M^{-1}B) \geq \lambda_i(M^{-1}A).
\end{aligned} \tag{6.3.4}$$

Proof. Note that under the assumption of the theorem we can prove that $(1 - \rho(M^{-1}D))^{-1} = 1 + \rho(B^{-1}D)$ as we did in the proof of Theorem 6.3.3. We now prove that

$$\lambda_i(M^{-1}B)\lambda_{\max}(B^{-1}M) \geq \lambda_i(M^{-1}A). \tag{6.3.5}$$

To this end, we need to prove that $\lambda_{\max}(B^{-1}A) \leq \lambda_{\max}(B^{-1}M)$. Equalities $A = B + R$ and $M = B + D$ show that $B^{-\frac{1}{2}}MB^{-\frac{1}{2}} - B^{-\frac{1}{2}}AB^{-\frac{1}{2}} = B^{-\frac{1}{2}}(D - R)B^{-\frac{1}{2}}$, which implies that

$$\lambda_{\max}(B^{-1}M) = \lambda_{\max}(B^{-\frac{1}{2}}MB^{-\frac{1}{2}}) \geq \lambda_{\max}(B^{-\frac{1}{2}}AB^{-\frac{1}{2}}) = \lambda_{\max}(B^{-1}A).$$

The inequality (6.3.5) now follows from the first inequality of (6.3.1)

Let $\widetilde{A} = \alpha I + A$, $\widetilde{B} = \alpha I + B$ and $\widetilde{M} = \alpha I + M$, where α is any positive number. One finds that $\widetilde{A} = \widetilde{B} + R$ and $\widetilde{M} = \widetilde{B} + D$. Applying (6.3.5) to $\widetilde{A}$, $\widetilde{B}$ and $\widetilde{M}$ shows that

$$\lambda_i(\widetilde{M}^{-1}\widetilde{B})\lambda_{\max}(\widetilde{B}^{-1}\widetilde{M}) \geq \lambda_i(\widetilde{M}^{-1}\widetilde{A}). \tag{6.3.6}$$

It follows from the proof of Theorem 6.3.1 that $\lambda_{\max}(\widetilde{B}^{-1}\widetilde{M}) = (1-\rho(\widetilde{M}^{-1}D))^{-1}$. Hence, we have from (6.3.6) that

$$(1-\rho(\widetilde{M}^{-1}D))^{-1}\lambda_i(\widetilde{M}^{-1}\widetilde{B}) \geq \lambda_i(\widetilde{M}^{-1}\widetilde{A}). \tag{6.3.7}$$

Using Courant-Fischer theorem (see e.g. [8]) and setting $\alpha \to 0$ in (6.3.7) show (6.3.4). □

If D is a diagonally compensative matrix of R, Theorem 6.3.2 (b) shows a bound of $\kappa(M^{-1}A)$. For the general case of compensative matrices the next result shows the same bound for matrix compensation if $B^{-1}R \geq 0$.

THEOREM 6.3.4. *Let $A = B+R$ and $M = B+D$, where A and B are positive definite and D is a compensative matrix of R. If $B^{-1}R \geq 0$ and $\rho(B^{-1}D) < 1$,*

$$\kappa(M^{-1}A) \leq \frac{1+\rho(B^{-1}D)}{1-\rho(B^{-1}D)}. \tag{6.3.8}$$

Proof. Since $D - R$ is positive semidefinite, $\lambda(B^{-1}R) \leq \rho(B^{-1}D)$. In the case of $B^{-1}R \geq 0$, Perron-Frobenius theorem (see e.g. [6]) shows that

$$\rho(B^{-1}R) = \lambda_{\max}(B^{-1}R) \leq \rho(B^{-1}D).$$

$0 < \lambda(M^{-1}A) \leq 1$ and (6.3.1) show that

$$\begin{aligned}
\kappa(M^{-1}A) &\leq \lambda_{\min}(M^{-1}A)^{-1} \leq (\lambda_{\min}(M^{-1}B)\lambda_{\min}(B^{-1}A))^{-1} \\
&= \frac{\lambda_{\max}(B^{-1}M)}{\lambda_{\min}(B^{-1}A)} = \frac{1+\lambda_{\max}(B^{-1}D)}{1+\lambda_{\min}(B^{-1}R)} \\
&\leq \frac{1+\rho(B^{-1}D)}{1-\rho(B^{-1}R)} \leq \frac{1+\rho(B^{-1}D)}{1-\rho(B^{-1}D)},
\end{aligned}$$

which is (6.3.8). □

As an application, we use diagonal compensation to estimate the spectral radius of positive definite matrices.

PROPOSITION 6.3.5. *Let $A = (a_{ij})_{i,j=1}^n = B + R$ be a positive definite matrix, where R is a nonnegative matrix and $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ is a diagonally compensative matrix of R. If all positive off-diagonal entries of A are reduced, then*

$$\rho(A) \leq 2(a+d) \tag{6.3.9}$$

where $a = \max(a_{11}, a_{22}, \ldots, a_{nn})$ and $d = \max(d_1, d_2, \ldots, d_n)$

Proof. Let $C = (c_{ij})_{i,j=1}^n$ be an M-matrix and $\mathbf{u} = (u_1, \ldots, u_n)^T$ be a positive vector such that $C\mathbf{u} > 0$. Then $U^{-1}CU$ is diagonally dominant, where $U = \mathrm{diag}(u_1, u_2, \ldots, u_n)$. Hence $\rho(C) = \rho(U^{-1}CU) \leq \|U^{-1}CU\|_\infty \leq 2c$, where $c = \max(c_{11}, c_{22}, \ldots, c_{nn})$. Under the assumption of the theorem, M defined by (6.2.2) becomes a Stieltjes matrix with the diagonal $(a_{11}+d_1, a_{22}+d_2, \ldots, a_{nn}+d_n)$. (6.3.9) follows from $\rho(A) \leq \rho(M)$. □

6.4. Applications to Preconditioners

Let M be a compensative matrix of A. Another way to construct a preconditioner for A is to choose a preconditioner of M as a preconditioner of A. To apply our results to the preconditioners, we first show the following result.

THEOREM 6.4.1. *Let A and M be positive definite such that $\|I-M^{-1}A\|_2 \le \varepsilon$, where $\varepsilon < 1$. Then for any positive matrix C*

$$\kappa(C^{-1}A) \le \frac{1+\varepsilon}{1-\varepsilon}\kappa(C^{-1}M).$$

Proof. Note that for any two positive definite matrices $\widetilde{A}$ and $\widetilde{B}$,

$$\rho(\widetilde{A}\widetilde{B}) = \rho(\widetilde{A}^{\frac{1}{2}}\widetilde{B}\widetilde{A}^{\frac{1}{2}}) \le \rho(\widetilde{A})\rho(\widetilde{B}). \tag{6.4.1}$$

Let $A - M = E$. Then $\rho(M^{-\frac{1}{2}}EM^{-\frac{1}{2}}) = \rho(I - M^{-1}A) \le \|I - M^{-1}A\|_2 \le \varepsilon < 1$, which implies that $I+M^{-\frac{1}{2}}EM^{-\frac{1}{2}}$ is positive definite. Applying (6.4.1) shows that

$$\begin{aligned}
&\lambda_{\max}(C^{-1}A) = \rho(C^{-\frac{1}{2}}AC^{-\frac{1}{2}}) = \rho(C^{-\frac{1}{2}}M^{\frac{1}{2}}(I + M^{-\frac{1}{2}}EM^{-\frac{1}{2}})M^{\frac{1}{2}}C^{-\frac{1}{2}}) \\
= \;& \rho(M^{\frac{1}{2}}C^{-1}M^{\frac{1}{2}}(I + M^{-\frac{1}{2}}EM^{-\frac{1}{2}})) \le \rho(M^{\frac{1}{2}}C^{-1}M^{\frac{1}{2}})\rho(I + M^{-\frac{1}{2}}EM^{-\frac{1}{2}}) \\
\le \;& \rho(C^{-1}M)(1 + \rho(M^{-\frac{1}{2}}EM^{-\frac{1}{2}})) \le \lambda_{\max}(C^{-1}M)(1+\varepsilon)
\end{aligned}$$

On the other hand, $\rho(M^{-\frac{1}{2}}EM^{-\frac{1}{2}}) < 1$ shows that

$$\begin{aligned}
A^{-1} &= (M+E)^{-1} = M^{-\frac{1}{2}}(I + M^{-\frac{1}{2}}EM^{-\frac{1}{2}})^{-1}M^{-\frac{1}{2}} \\
&= M^{-\frac{1}{2}}\Big(\sum_{k=0}^{\infty}(M^{-\frac{1}{2}}EM^{-\frac{1}{2}})^k\Big)M^{-\frac{1}{2}}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\lambda_{\max}(A^{-1}C) = \|C^{\frac{1}{2}}A^{-1}C^{\frac{1}{2}}\|_2 \\
\le \;& \|C^{\frac{1}{2}}M^{-1}C^{\frac{1}{2}}\|_2 \sum_{k=0}^{\infty}\|M^{-\frac{1}{2}}EM^{-\frac{1}{2}}\|_2^k \le \lambda_{\max}(M^{-1}C)\frac{1}{1-\varepsilon}.
\end{aligned}$$

The theorem follows immediately. □

A similar result for $C = M$ can be found in Axelsson's book [2] (pp. 327–328).

Let M be a compensative matrix of A and C be a preconditioner of M. If we choose C as a preconditioner of A, the following corollary shows a bound for $\kappa(C^{-1}A)$.

COROLLARY 6.4.2. *Let $A = B + R$ and $M = B + D$, Assume A and B are positive definite. If 1. D is a diagonally compensative matrix of R and $\rho(B^{-1}D) < 1$ or 2. D is a compensative matrix of R, $B^{-1}R \ge 0$ and $\rho(B^{-1}D) < 1$, then for any positive definite matrix C*

$$\kappa(C^{-1}A) \le \frac{1+\rho(B^{-1}D)}{1-\rho(B^{-1}D)}\kappa(C^{-1}M) \tag{6.4.2}$$

Proof. inequality (6.3.1) implies that $\kappa(C^{-1}A) \le \kappa(M^{-1}A)\,\kappa(C^{-1}M)$. Theorem 6.3.1 (b) and Theorem 6.3.4 complete (6.4.2). □

Acknowledgments

I am grateful to Professor O. Axelsson for valuable comments on the manuscript.

References

[1] O. AXELSSON, *A generalized SSOR method*, BIT, 13 (1972), pp. 443–467.

[2] ———, *Iterative Solution Methods*, Cambridge University Press, New York, 1994.

[3] O. AXELSSON AND V. BARKER, *Finite Element Solution of Boundary Value Problems*, Academic Press, Orlando, Fl , 1984.

[4] O. AXELSSON AND L. KOLOTILINA, *Diagonally compensated reduction and related preconditioning methods*, Numer. Linear Algebra Appl., 1 (1994), pp. 155–177

[5] R. BEAUWENS AND B. TOMBUYSES, *Graph perturbations*, Appl. Numer. Math., (to appear).

[6] A. BERMAN AND R. PLEMMONS, *Nonnegative Matrices in Mathematics Sciences*, Academic Press, New York, 1979.

[7] L. ELSNER AND V. MEHRMANN, *Convergence of block iterative methods for linear systems arising in the numerical solution of Euler equations*, Numer. Math., 59 (1991), pp. 541–559.

[8] G. GOLUB AND C. VAN LOAN, *Matrix Computation*, 2nd ed, The Johns Hopkins University Press, Baltimore, 1989.

[9] I GUSTAFSSON, *A class of first order factorization methods*, BIT, 18 (1978), pp. 142–156.

[10] R. NABBEN, *On a class of matrices which arise in the numerical solution of Euler equations*, Numer. Math., 63 (1992), pp. 411–431

CHAPTER 7

Eigenvalue Estimates for Incomplete Factorizations*

Abstract. Eigenvalue estimates of block incomplete preconditioners are considered. We investigate how the block diagonal entries and off block diagonal entries influence the bounds of all eigenvalues. The results presented here improve and unify some previous results. We generalize the well-known inequality that the spectral radius is bounded by the trace for symmetric positive semidefinite matrices to block form. Some of the methods can also be useful to estimate lower bounds of block incomplete preconditioners.

Key words. eigenvalue estimates, incomplete factorization, preconditioners

AMS subject classifications. 65F10, 65F15, 65F50

7.1. Introduction

To estimate the rate of convergence of preconditioned iterative methods such as the Chebyshev iterative method and the conjugate gradient iterative method, one needs to know the extreme eigenvalues and the distribution of eigenvalues of the preconditioned matrix respectively, see [1], [2], [8], [7], [5], [12]. Naturally, this problem by itself is difficult, especially for the distribution of all eigenvalues. Fortunately, it has been shown (see [4]) that under certain conditions lower and upper bounds of some eigenvalues can be derived and they provide the information necessary to compare modified and unmodified incomplete factorization methods for symmetric positive definite matrices, for instance.

Consider the implicit preconditioner on factorized form

$$C = (X + L)X^{-1}(X + L^T)$$

of a symmetric matrix A. Let $A = D_A + L_A + L_A^T$, where D_A is a block diagonal matrix. If A is a Stieltjes matrix and $L = L_A$ in some cases, some methods to estimate upper bounds of eigenvalues of $C^{-1}A$ were derived in [6], [4], [10], [9]. However, the assumptions limit the applicability of the results because for incomplete factorization methods they do not hold in general. In this paper, we discuss upper bounds and distribution of eigenvalues of block incomplete preconditioners for the general case of A being only a symmetric matrix. All of results allow that L_A differs from L. As we will see, even when the assumption of A is weakened, we can have strong results. The results here unify some of the previous results on upper bounds of eigenvalues of incomplete preconditioners.

---

* This chapter is based on the paper, O. Axelsson and H. Lu, *On eigenvalue estimates for block incomplete factorization methods*, SIAM J. Matrix Anal. Appl. 16 (1995) (to appear).

The paper is organized as follows: under assumption of A being a symmetric matrix, in §7.2 we focus our attention on both estimates of upper bounds and distribution of eigenvalues of block incomplete preconditioners. The result presented in this section can also be useful to estimate lower bounds of block incomplete preconditioners. In §7.3, some further useful methods to estimate upper bounds and distribution of eigenvalues are presented based on the fundamental result in the previous section. We generalize the well-known inequality $\rho(A) \leq \mathrm{tr}(A)$ for A symmetric positive semidefinite to block form, which with the result in §7.2 yields a new upper bound depending only on the block order of matrices for the largest eigenvalue.

For convenience, $\lambda_i(A)$ denotes the ith eigenvalue of matrix A and it is assumed that all eigenvalues of a matrix are ordered in a non-increasing order. For any pair of matrices A, B of the same order, $A \geq B$ means that the same inequality holds elementwise. The notation s.p.d. means symmetric positive definite while s.p.s.d. means symmetric positive semidefinite.

7.2. Upper and Lower Bounds of Eigenvalues

Let A be a symmetric matrix partitioned in a block form

$$A = D_A + L_A + L_A^T,$$

where D_A, L_A is the block diagonal part and strictly lower block triangular part of A, respectively. Consider a preconditioner C in the form

$$C = (X + L)X^{-1}(X + L^T),$$

i.e., a so called implicit preconditioner, where X is a block diagonal and s.p.d. matrix and L is a block lower triangular matrix. X and L are partitioned in blocks consistently with D_A and L_A, respectively. We present first a result for upper bounds of eigenvalues, which extends some results in [4], [10].

THEOREM 7.2.1. *Let A be symmetric and assume that X is s.p.d. and $\sigma X - K$ and $K - \beta X$ are s.p.s.d. for some constants σ, β. Then*

$$\lambda_i(M(\beta)) \leq \lambda_i(C^{-1}A) \leq \lambda_i(M(\sigma)), \tag{7.2.1}$$

where $K = A - L - L^T$, $C = (X + L)X^{-1}(X + L^T)$ and

$$M(\alpha) = (I + \tilde{L})^{-1} + (I + \tilde{L}^T)^{-1} + (\alpha - 2)(I + \tilde{L})^{-1}(I + \tilde{L}^T)^{-1}, \tag{7.2.2}$$
$$\tilde{L} = X^{-\frac{1}{2}} L X^{-\frac{1}{2}}.$$

Proof. We have

$$A = K - \sigma X + (X + L) + (X + L^T) + (\sigma - 2)X.$$

A computation with a similarity transformation of $C^{-1}A$ shows that

$$\begin{aligned}
& X^{-\frac{1}{2}}(X + L^T)C^{-1}A(X + L^T)^{-1}X^{\frac{1}{2}} \\
= \; & X^{\frac{1}{2}}(X + L)^{-1}A(X + L^T)^{-1}X^{\frac{1}{2}} \\
= \; & X^{\frac{1}{2}}(X + L)^{-1}(K - \sigma X)(X + L^T)^{-1}X^{\frac{1}{2}} \\
& + X^{\frac{1}{2}}(X + L)^{-1}X^{\frac{1}{2}} + X^{\frac{1}{2}}(X + L^T)^{-1}X^{\frac{1}{2}} \\
& + (\sigma - 2)X^{\frac{1}{2}}(X + L)^{-1}X(X + L^T)^{-1}X^{\frac{1}{2}} \\
= \; & X^{\frac{1}{2}}(X + L)^{-1}(K - \sigma X)(X + L^T)^{-1}X^{\frac{1}{2}} + M(\sigma).
\end{aligned}$$

Since, by assumption, $K - \sigma X$ is negative semidefinite, this shows that

$$\lambda_i(C^{-1}A) \leq \lambda_i(M(\sigma)).$$

Similarly, we can prove the first inequality in (7.2.1). □

If $L = L_A$ is non-positive and X is a block diagonal Stieltjes matrix, The special case of Theorem 7.2.1 for the maximum eigenvalue of $C^{-1}A$ can be found in [10].

The following five propositions give situations in which the theorem is applicable.

PROPOSITION 7.2.2. *X is s.p.s.d. if X is a symmetric Z-matrix and $X\mathbf{v} \geq 0$ for some vector $\mathbf{v} > 0$.*

Proof. Let $\mathbf{v} = (v_1, v_2, \ldots, v_k)^T$ and $D = \text{diag}(v_1, v_2, \ldots, v_k)$. $DXD + \varepsilon I$ is a diagonally dominant Z-matrix for any $\varepsilon > 0$, which implies, in particular, that X is s.p.s.d.. □

PROPOSITION 7.2.3. *Let X be symmetric. If $\sigma X - D_A$ is a Z-matrix and the entries of $L + L^T$ are not larger than the corresponding entries of $L_A + L_A^T$, then $\sigma X - K$ is a Z-matrix. If, in addition, $\sigma X\mathbf{v} - K\mathbf{v} \geq 0$ for some positive vector $\mathbf{v}$, then $\sigma X - K$ is s.p.s.d..*

Proof. A direct calculation shows that $\sigma X - K = (\sigma X - D_A) + (L + L^T - L_A - L_A^T)$, which shows that $\sigma X - K$ is a Z-matrix. An application of Proposition 7.2.2 completes the proof. □

PROPOSITION 7.2.4. *$\lambda_i(M(\sigma)) \leq \min(\lambda_i(M(2)), 1/(2-\sigma))$ if $\sigma \in [0, 2]$. The inequality is strict if $\sigma < 2$.*

Proof. $\lambda_i(M(\sigma)) \leq \lambda_i(M(2))$ is straightforward. By a simple computation

$$\begin{aligned} M(\sigma) &= (I + \tilde{L})^{-1} + (I + \tilde{L}^T)^{-1} + (\sigma - 2)(I + \tilde{L})^{-1}(I + \tilde{L}^T)^{-1} \\ &= \frac{1}{2-\sigma} I - (2-\sigma)((I + \tilde{L})^{-1} - \frac{1}{2-\sigma} I)((I + \tilde{L}^T)^{-1} - \frac{1}{2-\sigma} I), \end{aligned}$$

we finish the proof. □

If $L = L_A$, the upper bound $1/(2-\sigma)$ can be found in [4].

PROPOSITION 7.2.5. *Suppose matrix X is s.p.d. and $\sigma X - K + \gamma I$ is s.p.s.d. Then $(\sigma + \gamma/\lambda_{\min}(X))X - K$ is s.p.s.d. if $\gamma \geq 0$ and $(\sigma + \gamma/\lambda_{\max}(X))X - K$ is s.p.s.d. if $\gamma \leq 0$.*

PROPOSITION 7.2.6. *$\lambda_{\max}(X^{-1}K)X - K$ and $\lambda_{\min}(X^{-1}K)X + K$ are s.p.s.d. if X is s.p.d..*

7.3. Some Alternative Upper Bounds

As we have seen in the previous section, the maximum eigenvalue of $C^{-1}A$ can be bounded by $\frac{1}{2-\sigma}$ if $\sigma < 2$, but the situation is not so fortunate if $\sigma > 2$. It is impossible to derive a bound by involving σ alone. The bound of the eigenvalues must depend on both σ and the lower triangular matrix $\tilde{L}$. In this section, first, we discuss how to estimate the eigenvalue bound of $C^{-1}A$ if $\sigma > 2$. Though $\frac{1}{2-\sigma}$ is an upper bound provided $\sigma < 2$, it is still very large if σ is close to 2. In the second half of this section, we reconsider the bound in this case. The discussion is based on our generalization of the well-known inequality $\rho(A) \leq \text{tr}(A)$ for A s.p.s.d. to

block form. It is shown that $2-\sigma+2(\sigma-1)m$ is another upper bound of $\lambda(C^{-1}A)$ if $1<\sigma\leq 2$ and A is an $m\times m$ block s.p.s.d. matrix.

Let $\tilde{M}=(I+\tilde{L})(I+\tilde{L}^T)$, where $\tilde{L}$ stands for the same matrix as in Theorem 7.2.1. The following result gives a method to estimate upper bounds of eigenvalues $\lambda_i(C^{-1}A)$ if $\sigma\geq 2-\lambda_{n-i+1}^{\frac{1}{2}}(\tilde{M})=2-\lambda_i^{-\frac{1}{2}}(\tilde{M}^{-1})$.

THEOREM 7.3.1. *Let matrices A and C satisfy the conditions of* Theorem 7.2.1. *If $\kappa_i\geq\lambda_i(\tilde{M}^{-1})$, where $\tilde{M}=(I+\tilde{L})(I+\tilde{L}^T)$, and $\sigma\geq 2-\kappa_i^{-\frac{1}{2}}$, then*

$$\lambda_i(C^{-1}A)\leq(\sigma-2)\kappa_i+\kappa_i^{\frac{1}{2}}. \tag{7.3.1}$$

Proof. Using Weyl's theorem (cf. Parlett [11], p. 192), we find for any $\mu<2$ and $\mu\leq\sigma$ that

$$\begin{aligned}\lambda_i(M(\sigma)) &= \lambda_i((I+\tilde{L})^{-1}+(I+\tilde{L}^T)^{-1}+(\sigma-2)(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1})\\ &\leq \lambda_{\max}((I+\tilde{L})^{-1}+(I+\tilde{L}^T)^{-1}+(\mu-2)(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1})\\ &\quad+(\sigma-\mu)\lambda_i(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1})\\ &\leq \frac{1}{2-\mu}+(\sigma-\mu)\kappa_i.\end{aligned}$$

Therefore, $\lambda_i(M(\sigma))\leq\min_{\mu<2}((2-\mu)^{-1}+(\sigma-\mu)\kappa_i=2\kappa_i^{\frac{1}{2}}+(\sigma-2)\kappa_i$. The minimum is taken for $\mu=2-\kappa_i^{-\frac{1}{2}}$. An application of Theorem 7.2.1 ends the proof of inequality (7 3.1). □

The bound given by (7.3.1) is clearly an improvement of $1/(2-\sigma)$ if $\sigma\geq 2-\lambda_{n-i+1}^{-\frac{1}{2}}(\tilde{M})$.

Divide $I+\tilde{L}$ into a 2×2 block form

$$I+\tilde{L}=\begin{pmatrix}\tilde{L}_1 & 0\\ \tilde{L}_{21} & \tilde{L}_2\end{pmatrix},$$

where $\tilde{L}_1$ and $\tilde{L}_2$ are the $k\times k$ and $(m-k)\times(m-k)$ block lower triangular submatrices of $I+\tilde{L}$, respectively, and $\tilde{L}_{21}$ is the $(m-k)\times k$ block submatrix of $I+\tilde{L}$ at the south-west corner.

Let $\Lambda(D)$ denote the set of all nonzero eigenvalues of matrix D. Let G_1 be an $n\times m$ matrix and G_2 is an $m\times n$ matrix. It is well known that $\Lambda(G_1G_2)=\Lambda(G_2G_1)$ Set $\tilde{M}_i=\tilde{L}_i\tilde{L}_i^T$, $i=1,2$, and denote $\tilde{k}=n_1+\ldots+n_k$, $\tilde{p}=n_{k+1}+\ldots+n_m$. We now consider how to estimate $\lambda_i(\tilde{M})$. To this end, we need the following lemmas:

LEMMA 7.3.2. *Let α_1 and α_2 be positive numbers, B be a $p\times k$ real matrix, and D be a matrix of the form*

$$D=\begin{pmatrix}\alpha_1^{\frac{1}{2}}I_k & 0\\ B & \alpha_2^{\frac{1}{2}}I_p\end{pmatrix}\begin{pmatrix}\alpha_1^{\frac{1}{2}}I_k & B^T\\ 0 & \alpha_2^{\frac{1}{2}}I_p\end{pmatrix}, \tag{7.3.2}$$

where I_k denotes the unit matrix of order k. Then the eigenvalues of D are given by

$$\lambda_i(D)=\begin{cases}f_+(\alpha_1,\alpha_2,\mu_i), & \text{if } 1\leq i\leq\min(p,k)\\ \alpha_1, & \text{if } p<i\leq k\\ \alpha_2, & \text{if } k<i<\leq p\\ f_-(\alpha_1,\alpha_2,\mu_{k+p+1-i}), & \text{if } \max(p,k)<i\leq k+p,\end{cases}$$

where $\mu_1, \ldots, \mu_{\min(p,k)}$ *are the first* $\min(p,k)$ *eigenvalues of* BB^T, $f_+(\alpha,\beta,\mu)$ *and* $f_-(\alpha,\beta,\mu)$ *are the largest resp the smallest zero of the function*

$$t \rightsquigarrow (\alpha - t)(\beta - t) - \mu t.$$

Proof. If $k \geq p$, a computation, using a block decomposition of $\lambda I_{k+p} - D$ shows the characteristic polynomial of D

$$\begin{aligned} f_D(\lambda) &= \det(\lambda I_{k+p} - D) \\ &= (\lambda - \alpha_1)^{k-p} \det((\lambda - \alpha_1)(\lambda - \alpha_2) I_p - \lambda BB^T) \\ &= (\lambda - \alpha_1)^{k-p} \prod_{i=1}^{p} ((\lambda - \alpha_1)(\lambda - \alpha_2) - \mu_i \lambda). \end{aligned}$$

Thus, $f_+(\alpha_1,\alpha_2,\mu_i)$, $f_-(\alpha_1,\alpha_2,\mu_i)$ and α_1 are eigenvalues of D. Since $f_+(\alpha_1,\alpha_2,x)$ and $f_-(\alpha_1,\alpha_2,x)$ are monotonously increasing and monotonously decreasing, respectively, the lemma follows immediately due to the fact that $f_+(\alpha_1,\alpha_2,x) \geq \max(\alpha_1,\alpha_2)$ and $f_-(\alpha_1,\alpha_2,x) \leq \min(\alpha_1,\alpha_2)$ for $x \geq 0$. Note that $\Lambda(B^TB) = \Lambda(BB^T)$. Similarly, one can prove the case $k < p$. □

LEMMA 7.3.3. *With* $L = \begin{pmatrix} L_1 & 0 \\ L_{21} & L_2 \end{pmatrix}$, *where* L_i *are nonsingular, and* $\tilde{L} = \begin{pmatrix} \sigma_1 & 0 \\ L_{21} & \sigma_2 \end{pmatrix}$, *where* $\sigma_i^2 = \lambda_{\min}(L_i^T L_i)$, *we have that* $\lambda_i(LL^T) \geq \lambda_i(\tilde{L}\tilde{L}^T)$.

Proof. Note that it is readily seen that $\lambda_i(BG) \geq \lambda_i(DG)$ if B, D, G and $B - D$ are s.p.s.d.. Hence, we have

$$\begin{aligned} \lambda_i(LL^T) &= \lambda_i\left(\begin{pmatrix} L_1 & 0 \\ 0 & I_p \end{pmatrix} \begin{pmatrix} I_k & 0 \\ L_{21} & L_2 \end{pmatrix} \begin{pmatrix} I_k & L_{21}^T \\ 0 & L_2^T \end{pmatrix} \begin{pmatrix} L_1^T & 0 \\ 0 & I_p \end{pmatrix} \right) \\ &= \lambda_i\left(\begin{pmatrix} L_1^T L_1 & 0 \\ 0 & I_p \end{pmatrix} \begin{pmatrix} I_k & 0 \\ L_{21} & L_2 \end{pmatrix} \begin{pmatrix} I_k & L_{21}^T \\ 0 & L_2^T \end{pmatrix} \right) \\ &\geq \lambda_i\left(\begin{pmatrix} \sigma_1^2 I_k & 0 \\ 0 & I_p \end{pmatrix} \begin{pmatrix} I_k & 0 \\ L_{21} & L_2 \end{pmatrix} \begin{pmatrix} I_k & L_{21}^T \\ 0 & L_2^T \end{pmatrix} \right) \\ &= \lambda_i\left(\begin{pmatrix} \sigma_1 I_k & 0 \\ 0 & I_p \end{pmatrix} \begin{pmatrix} I_k & 0 \\ L_{21} & L_2 \end{pmatrix} \begin{pmatrix} I_k & L_{21}^T \\ 0 & L_2^T \end{pmatrix} \begin{pmatrix} \sigma_1 I_k & 0 \\ 0 & I_p \end{pmatrix} \right) \\ &= \lambda_i\left(\begin{pmatrix} \sigma_1 I_k & 0 \\ L_{21} & L_2 \end{pmatrix} \begin{pmatrix} \sigma_1 I_k & L_{21}^T \\ 0 & L_2^T \end{pmatrix} \right) \equiv \tilde{\lambda}_i. \end{aligned}$$

Similarly, it follows that $\tilde{\lambda}_i \geq \lambda_i(\tilde{L}\tilde{L}^T)$. □

Since $\Lambda(\tilde{L}_{21}\tilde{L}_{21}^T) = \Lambda(\tilde{L}_{21}^T\tilde{L}_{21})$, it follows from the proof of Theorem 7.3.1 with using Lemma 7.3.2 and Lemma 7.3.3 that

$$\lambda_i(\tilde{M}) \geq \begin{cases} f_+(\lambda_{min}(\tilde{M}_1), \lambda_{min}(\tilde{M}_2), \mu_i), & \text{if } 1 \leq i \leq \min(\tilde{p},\tilde{k}) \\ \lambda_{min}(\tilde{M}_1), & \text{if } \tilde{p} < i \leq \tilde{k} \\ \lambda_{min}(\tilde{M}_2), & \text{if } \tilde{k} < i \leq \tilde{p} \\ f_-(\lambda_{min}(\tilde{M}_1), \lambda_{min}(\tilde{M}_2), \mu_{n+1-i}), & \text{if } \max(\tilde{p},\tilde{k}) < i \leq n \end{cases}$$

where $\mu_1, \mu_2, \ldots, \mu_{\min(\tilde{p},\tilde{k})}$ are the first $\min(\tilde{p},\tilde{k})$ eigenvalues of $L_{21}L_{21}^T$ numbered in a non-increasing order.

LEMMA 7.3.4. *Let B be a $p \times k$ real matrix and D be a matrix of the form*

$$D = \begin{pmatrix} \alpha_1 I_k & B^T \\ B & \alpha_2 I_p \end{pmatrix}.$$

Then the eigenvalues of D are given by

$$\lambda_i(D) = \begin{cases} g_+(\alpha_1, \alpha_2, \beta_i), & \text{if } 1 \le i \le \min(p,k) \\ \alpha_1, & \text{if } p < i \le k \\ \alpha_2, & \text{if } k < i \le p \\ g_-(\alpha_1, \alpha_2, \beta_{k+p+1-i}), & \text{if } \max(p,k) < i \le k+p, \end{cases}$$

where $\beta_1, \beta_2, \ldots,$ and $\beta_{\min(p,k)}$ are the first $\min(p,k)$ eigenvalues of BB^T,

$$g_+(\alpha_1, \alpha_2, x) = \frac{1}{2}(\alpha_1 + \alpha_2 + ((\alpha_1 - \alpha_2)^2 + 4x)^{\frac{1}{2}}),$$

$$g_-(\alpha_1, \alpha_2, x) = \frac{1}{2}(\alpha_1 + \alpha_2 - ((\alpha_1 - \alpha_2)^2 + 4x)^{\frac{1}{2}}).$$

Proof. The proof is similar to that of Lemma 7.3.2. □

THEOREM 7.3.5. *Let $A = (A_{ij})_{i,j=1}^m$ be a block matrix partitioning of a s.p.s.d. matrix. Then*

$$\rho(A) \le \sum_{i=1}^{m} \rho(A_{ii}). \tag{7.3.3}$$

Proof. Consider first the case $m = 2$. Let $B = \begin{pmatrix} \rho(A_{11})I & A_{12} \\ A_{21} & \rho(A_{22})I \end{pmatrix}$ which is clearly s.p.s.d. and $\rho(A) \le \rho(B)$. Lemma 7.3.4 shows that

$$\rho(A_{11}) + \rho(A_{22}) - ((\rho(A_{11}) - \rho(A_{22})^2 + \|A_{12}\|_2^2)^{\frac{1}{2}} = 2\lambda_{\min}(B) \ge 0$$

and, hence

$$\rho(B) \le \frac{1}{2}(\rho(A_{11}) + \rho(A_{22}) + ((\rho(A_{11}) - \rho(A_{22})^2 + \|A_{12}\|_2^2)^{\frac{1}{2}}) \le \rho(A_{11}) + \rho(A_{22}).$$

By induction, we find that for any s.p.s.d. matrix, $A = (A_{ij})_{i,j=1}^m$ (7.3.3) holds. □

If A has scalar form, (7.3.3) reduces to the well-known inequality

$$\rho(A) \le \mathrm{tr}(A).$$

If $1 < \sigma \le 2$, we give now an alternative method to estimate an upper bound of the eigenvalues of $C^{-1}A$, which yields $2 - \sigma + 2(\sigma - 1)m$ as an upper bound if A is s.p.s.d..

THEOREM 7.3.6. *Let A and C satisfy all conditions of* Theorem 7.2.1. *If A s.p.s.d. and $1 < \sigma \le 2$, then*

$$\lambda_{max}(C^{-1}A) \le 2 - \sigma + 2(\sigma - 1)m.$$

Proof.

$$(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1} = (\sum_{i=0}^{m-1}(-1)^i\tilde{L}^i)(\sum_{i=0}^{m-1}(-1)^i(\tilde{L}^T)^i)$$
$$= I + \sum_{i=1}^{m-1}(-1)^i\tilde{L}^i + \sum_{i=1}^{m-1}(-1)^i(\tilde{L}^T)^i) + (\sum_{i=1}^{m-1}(-1)^i\tilde{L}^i)(\sum_{i=1}^{m-1}(-1)^i(\tilde{L}^T)^i)$$
$$= M(2) - I + (\sum_{i=1}^{m-1}(-1)^i\tilde{L}^i)(\sum_{i=1}^{m-1}(-1)^i(\tilde{L}^T)^i).$$

Since $(\sum_{i=1}^{m-1}(-1)^i\tilde{L}^i)(\sum_{i=1}^{m-1}(-1)^i(\tilde{L}^T)^i)$ is s.p.s.d. and $1 < \sigma \leq 2$, Theorem 7.2.1 and the above yield

$$\lambda_i(C^{-1}A) \leq \lambda_i(M(2) + (\sigma-2)(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1})$$
$$= \lambda_i((2-\sigma)I + (\sigma-1)M(2) + (\sigma-2)(\sum_{i=1}^{m-1}(-1)^i\tilde{L}^i)(\sum_{i=1}^{m-1}(-1)^i(\tilde{L}^T)^i)$$
$$\leq 2 - \sigma + (\sigma-1)\lambda_i(M(2)).$$

Since A is s.p.s.d and C is s p.d., Proposition 7.2.4 shows that $M(2)$ is s.p s.d.. Since the diagonal part of $M(2)$ is $2I$, (7.3 3) finishes the proof. □

Under some additional assumptions, the bound of Theorem 7.3.6 can be reduced. For example, if

1. A is a Stieltjes matrix,
2. $L = L_A$,
3. X is a Stieltjes matrix such that offdiag$(X) \leq$offdiag(D_A), and
4. there is a positive vector $\mathbf{v}$ such that

$$C\mathbf{v} \geq 0, \tag{7.3.4}$$
$$(L^T - L)\mathbf{v} \leq C\mathbf{v}, \tag{7.3.5}$$
$$(X + L^T)\mathbf{v} \geq A\mathbf{v}, \tag{7.3 6}$$
$$L_p^T\mathbf{v} > 0,$$

where $X = L_p P L_p^T$ denotes the point LU decomposition of X, then it has been shown in [9] that

$$\lambda_{\max}(C^{-1}A) \leq m + 1.$$

These assumptions imply actually that $2X + L + L^T - A$ is s.p.s.d. In this case, $2X + L + L^T - A$ is a Stieltjes matrix. This follows, because as has been shown in [9], (7.3.4) and (7.3.5) imply $(X + L)\mathbf{v} \geq 0$. Hence using (7.3.6) shows that

$$(2X + L + L^T - A)\mathbf{v} \geq 0,$$

which implies that $2X + L + L^T - A$ is s p.s.d..

The matrix $M(2)$ acts as a key for estimating the eigenvalues $\lambda_i(C^{-1}A)$ as we have seen. In general, $\lambda_i(M(2))$ can be estimated by using Lemma 7.3.4. Denote $\bar{M}_i = \tilde{L}_i^{-1} + (\tilde{L}_i^T)^{-1}$, $i = 1, 2$, and $\eta_i = \lambda_{\max}(\bar{M}_i)$. According to the partitioning of $I + \tilde{L}$, it is easy to check that

$$M(2) = \begin{pmatrix} \bar{M}_1 & \bar{L}_{21}^T \\ \bar{L}_{21} & \bar{M}_2 \end{pmatrix}$$

and hence

$$\lambda_i(M(2)) \le \lambda_i \begin{pmatrix} \eta_1 I_{\tilde{k}} & \bar{L}_{21}^T \\ \bar{L}_{21} & \eta_2 I_{\tilde{p}} \end{pmatrix}.$$

Again using $\Lambda(\bar{L}_{21}\bar{L}_{21}^T) = \Lambda(\bar{L}_{21}^T\bar{L}_{21})$ and Lemma 7.3.4, we have that

$$\lambda_i(M(2)) \le \begin{cases} g_+(\eta_1, \eta_2, \gamma_i), & \text{if } 1 \le i \le \min(\tilde{p}, \tilde{k}) \\ \eta_1, & \text{if } \tilde{p} < i \le \tilde{k} \\ \eta_2, & \text{if } \tilde{k} < i \le \tilde{p} \\ g_-(\eta_1, \eta_2, \gamma_{n+1-i}), & \text{if } \max(\tilde{p}, \tilde{k}) < i \le n \end{cases}$$

where $\bar{L}_{21} = -\tilde{L}_1^{-1}\tilde{L}_{21}\tilde{L}_2^{-1}$, $\gamma_1, \ldots, \gamma_{\min(\tilde{p},\tilde{k})}$ are the first $\min(\tilde{p}, \tilde{k})$ eigenvalues of $\bar{L}_{21}\bar{L}_{21}^T$ numbered in a non-increasing order.

7.4. Application to Generalized SSOR Preconditioned Matrices

As an application of the results presented in §7.2 and §7.3, we now consider upper bounds of the condition number of the preconditioned matrix when the generalized SSOR method is applied to symmetric block tridiagonal matrices.

Let A be a block tridiagonal matrix of the form

$$A = \text{blocktridiag}(A_{i,i-1}, A_{ii}, A_{i,i+1}),$$

where $A_{ii} = \text{tridiag}(-b, a, -b)$ and $A_{i,i-1} = A_{i,i+1} = -cI$, $i = 1, 2, \ldots, m$. All blocks have order $n \times n$. In addition, we assume that $b, c \ge 0$ and $a \ge 2(b + c)$ Consider

$$A = D_A - L - L,$$

a splitting of A, and the generalized SSOR preconditioned matrix

$$C = (D - L)D^{-1}(D - L^T),$$

where $D_A = \text{blockdiag}(A_{11}, A_{22}, \ldots, A_{mm})$, L is the lower blocktridiagonal part of A, $D = \text{blockdiag}(D_1, D_2, \ldots, D_m)$ partitioned as D_A.

We compute a preconditioner C for A in a common way as follows (see [3]):

$$D_1 = A_{11},$$
$$D_i = A_{ii} - A_{i,i-1}X_{i-1}A_{i-1,i} + D_i', \qquad i = 2, 3, \ldots, m,$$

where X_i, $i \ge 1$, is a sparse approximation to D_i^{-1} and D_i' is a diagonal matrix such that

$$D_i'v = A_{i,i-1}(X_{i-1} - D_{i-1}^{-1})A_{i-1,i}v, \qquad i = 2, 3, \ldots, m$$

for some positive vector v. Hence we have

$$D_1 v = A_{11}v, \tag{7.4.1}$$

$$D_i v = (A_{ii} - A_{i,i-1}D_{i-1}^{-1}A_{i-1,i})v, \qquad i = 2, 3, \ldots, m. \tag{7.4.2}$$

Since $A_{ii} = (a-2b)I + b\,\text{tridiag}(-1, 2, -1)$, the smallest eigenvalue of A_{ii} is $a - 2b + b(2\sin\frac{\pi}{2(n+1)})^2$, denoted by λ, where n is the order of A_{ii}. Let $\mathbf{v}$ be the eigenvector of A_{ii} corresponding to λ. (7.4.1) and (7.4.2) imply that $\mathbf{v}$ is also an eigenvector of D_i and the corresponding smallest eigenvalue of D_i becomes

$$\lambda_1 = \lambda, \qquad \lambda_i = \lambda - c^2\lambda_{i-1}^{-1}, \qquad i = 2, 3, \ldots, m.$$

It is readily seen that λ_i converges monotonically to the lower bound

$$\tilde{\lambda} = \frac{1}{2}(\lambda + (\lambda^2 - 4c^2)^{\frac{1}{2}}).$$

Let $\sigma = \frac{\lambda}{\lambda - c^2\tilde{\lambda}^{-1}}$. We have $\sigma = 1 + \frac{c^2}{\lambda\tilde{\lambda} - c^2} < 2$. A computation shows that

$$\sigma D_1 \mathbf{v} - A_{11}\mathbf{v} \geq 0,$$

$$\sigma D_i \mathbf{v} - A_{ii}\mathbf{v} = \lambda \frac{\lambda - c^2\lambda_{i-1}^{-1}}{\lambda - c^2\tilde{\lambda}^{-1}}\mathbf{v} - \lambda\mathbf{v} \geq 0, \qquad i = 2, 3, \ldots, m,$$

which implies that $\sigma X - L - L^T - A$ is s.p.s.d.. Now Proposition 7.2.4 shows that

$$\rho(C^{-1}A) \leq \frac{1}{2 - \sigma} = \frac{\lambda\tilde{\lambda} - c^2}{\lambda\tilde{\lambda} - 2c^2}.$$

On the other hand, using $\lambda \geq 2c$, it is also seen that $\lambda_i \geq \hat{\lambda}_i$, which satisfy

$$\hat{\lambda}_1 = 2c, \ \hat{\lambda}_i = 2c - c^2\hat{\lambda}_i^{-1}, \ i = 2, 3, \ldots, m.$$

Hence $\hat{\lambda}_i = \frac{i+1}{i}c$. Let $\alpha = \frac{\hat{\lambda}_1}{\hat{\lambda}_1 - c^2\hat{\lambda}_m^{-1}} = 2 - \frac{2}{m+2}$. Similarly, we find that $\alpha D - L - L^T - A$ is also s p.s.d..

Consider the lower block tridiagonal matrix $T = I - D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$. We have

$$T^{-1} = (T_{ij}) = \sum_{t=0}^{m-1} (D^{-\frac{1}{2}}LD^{-\frac{1}{2}})^t,$$

where $T_{ii} = I_n$, $T_{ij} = c^{i-j}D_i^{-\frac{1}{2}}D_{i-1}^{-1}\ldots D_{j+1}^{-1}D_j^{-\frac{1}{2}}$, $i > j$, $T_{ij} = 0$, $i < j$.

Partition $(T^T)^{-1}T^{-1}$ into an $m \times m$ block matrix (B_{ij}) consistently with the partitioning of A. Clearly, $(T^T)^{-1}T^{-1}$ is a nonnegative matrix. Applying $D_i^{-1}\mathbf{v} \leq \frac{1}{(i+1)c}\mathbf{v}$ shows that

$$T_{ij}\mathbf{v} \leq \left(\frac{j(j+1)}{i(i+1)}\right)^{\frac{1}{2}}\mathbf{v} \leq \frac{j+1}{i+1}\mathbf{v} \text{ and } T_{ij}^T\mathbf{v} \leq \frac{j+1}{i+1}\mathbf{v}, \ i \geq j.$$

Hence,

$$B_{ij}\mathbf{v} = \sum_{k=1}^{m} T_{ki}^T T_{kj}\mathbf{v} \leq \sum_{k \geq \max(i,j)}^{m} \frac{(i+1)(j+1)}{(k+1)^2}\mathbf{v}$$

$$\sum_{j=1}^{m} B_{ij}\mathbf{v} = \sum_{j=1}^{i-1}\sum_{k=i}^{m} \frac{(i+1)(j+1)}{(k+1)^2}\mathbf{v} + \sum_{j=i}^{m}\sum_{k=j}^{m} \frac{(i+1)(j+1)}{(k+1)^2}\mathbf{v}$$

$$= \sum_{k=i}^{m}\sum_{j=1}^{k} \frac{(i+1)(j+1)}{(k+1)^2}\mathbf{v} = \frac{i+1}{2}\sum_{k=i}^{m}\left(1 + \frac{1}{k+1} - \frac{2}{(k+1)^2}\right)\mathbf{v}$$

$$< \frac{i+1}{2}\sum_{k=i}^{m}\left(1 + \frac{1}{k+1} - \frac{2}{(k+1)(k+2)}\right)\mathbf{v}$$

$$\leq \frac{i+1}{2}\left(m - i + 1 + \int_{i+1}^{m+1}\frac{dx}{x} + \frac{1}{i+1} - 2\sum_{k=i}^{m}\left(\frac{1}{k+1} - \frac{1}{k+2}\right)\right)\mathbf{v}$$

$$= \frac{1}{2}\left((i+1)\left(m - i + 1 + \log(m+1) - \log(i+1) + \frac{2}{m+2}\right) - 1\right)\mathbf{v}$$

$$< \frac{1}{8}\left(\left(m + 1 + \log(2) + \frac{2}{m+2}\right)\left(m + 3 + \log(2) + \frac{2}{m+2}\right) - 4\right)\mathbf{v}$$

$$< \frac{1}{8}(m + 2 + \log(2))^2\mathbf{v},$$

which implies that

$$\rho(\tilde{M}^{-1}) = \rho((T^T)^{-1}T^{-1}) \leq \frac{1}{8}(m+2+\log(2))^2 \equiv \kappa,$$

where $\tilde{M}$ stands for the same matrix as in Theorem 7.3.1. Since $\kappa \geq \rho(\tilde{M}^{-1})$ and $\alpha > 2 - \kappa^{-\frac{1}{2}}$, applying Theorem 7.3.1 shows that

$$\rho(C^{-1}A) \leq (\alpha-2)\kappa + 2\kappa^{\frac{1}{2}} \leq \frac{2\sqrt{2}-1}{4}(m+2+\log(2)).$$

This bound is approximately $0.4571(m+2+\log(2))$. The result can be further improved if we can estimate $\rho(\tilde{M}^{-1})$ more accurately. Application of the result in [9] (Theorem 4.3) shows only an $m/2$ upper bound for this example, although the result requires that A is a Stieltjes matrix, $L = L_A$ and some other additional conditions.

Further, because $A-C$ is a Z-matrix and $(A-C)\mathbf{v} = 0$, we have $\lambda_{\min}(C^{-1}A) = 1$ and, therefore,

$$\mathrm{cond}(C^{-1}A) \leq \min\left(\frac{\lambda\tilde{\lambda}-c^2}{\lambda\tilde{\lambda}-2c^2}, \frac{2\sqrt{2}-1}{4}(m+2+\log(2))\right).$$

For the model second order elliptic difference equation on a rectangular $n \times m$ mesh with uniform meshwidth $h = \frac{1}{n+1}$, we have $A_{ii} = \mathrm{tridiag}(-1,4,-1)$, $c = 1$. In this case, using the previously given bound on λ, we find

$$\frac{\lambda\tilde{\lambda}-1}{\lambda\tilde{\lambda}-2} \simeq \frac{n+1}{2\pi}.$$

Therefore,

$$\mathrm{cond}(C^{-1}A) \leq \min\left(\frac{n+1}{2\pi}, \frac{2\sqrt{2}-1}{4}(m+2+\log(2))\right).$$

It turns out that the second part holds also for the more common choice of the vector $\mathbf{e} = (1,1,\ldots,1)^T$, because we have $\tilde{D}_i\mathbf{e} \geq \frac{i+1}{i}\mathbf{e}$, where $\tilde{D}_i$ are the corresponding matrices of D_i obtained by using $\mathbf{e}$.

The general bound $2m$ of the condition number is hence not very accurate for the model type problem. Bounds involving only m, the number of blocks, are of particular interest when an elliptic second order difference equation is solved on an oblong rectangular domain with number of nodepoints $N_1 \times N_2$ where we assume that $N_1 \gg N_2$. If we number the points such that the order of the matrix blocks is N_1, i.e., there are $m = N_2$ blocks in the main diagonal, then applying Theorem 7.3.6 shows that

$$\mathrm{cond}(C^{-1}A) \leq 2N_2,$$

or $0.4571(N_2+2+\log(2))$ for the model problem, both of which hence do not depend on N_1. It is therefore efficient to choose big blocks for such domains.

Acknowledgments

Careful comments by two anonymous referees regarding the presentation of some results are gratefully acknowledged.

References

[1] O. AXELSSON, *A class of iterative methods for finite element equations*, Comput. Methods Appl. Mech Engrg , 9 (1976), pp. 123–137.

[2] ———, *Solution of linear systems of equations: iterative methods*, in Sparse Matrix Techniques, Lecture Notes in Mathematics 572, V. A. Barker, ed., Springer Verlag, Berlin, Heidelberg, New York, 1977, pp. 1–50.

[3] ———, *Incomplete block-matrix factorization preconditioning methods. the ultimate answer?*, J. Comp. Appl. Math., 12 & 13 (1985), pp 3–18.

[4] ———, *Bounds of eigenvalues of preconditioned matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 847–862.

[5] O. AXELSSON AND G. LINDSKOG, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 48 (1986), pp. 499–523.

[6] R. BEAUWENS, *Upper eigenvalue bounds for pencils of matrices*, Linear Algebra Appl., 62 (1984), pp. 87–104.

[7] A. GREENBAUM, *Comparison of splittings used with the conjugate gradient algorithm*, Numer. Math , 33 (1979), pp. 181–194

[8] A. JENNINGS, *Influence of the eigenvalues spectrum of the convergence rate of the conjugate gradient method*, IMA Journal of Numerical Analysis, 20 (1977), pp. 61–72.

[9] M M. MAGOLU, *Analytical bounds for block approximate factorization methods*, Linear Algebra Appl , 179 (1993), pp. 33–57.

[10] M.M. MAGOLU AND Y. NOTAY, *On the conditioning analysis of block approximate factorization methods*, Linear Algebra Appl., 154-156 (1991), pp. 583–599.

[11] B.N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, N.J., 1980.

[12] A. VAN DER SLUIS AND H A. VAN DER VORST, *The rate of convergence conjugate gradients*, Numer. Math. 48 (1986), pp. 543–560.

CHAPTER 8

Conditioning Analysis and Its Application*

Abstract. The paper deals with eigenvalue estimates for block incomplete factorization methods for symmetric matrices. Some previous results on upper bounds of maximum eigenvalue of preconditioned matrices are generalized to every eigenvalue. Considering the relation of matrices with their preconditioners, we derive an accurate upper bound depending only on the block order. Finally, the results are used to estimate the bound for every eigenvalue of the preconditioned matrices if the generalized SSOR method is used to solve an elliptic equation in two dimensions. Using a transformation we show how the coefficients of the differential equations influence the bounds of the eigenvalues of the preconditioned matrix.

Key words. eigenvalue estimates, incomplete factorization, preconditioners, elliptic equation

AMS subject classifications. 65F10, 65F15, 65F50

8.1. Introduction

Consider block incomplete factorizations of the form

$$C = (X + L)X^{-1}(X + L^T)$$

for symmetric matrices, where X is symmetric positive definite or, briefly s.p.d. and L is a lower triangular matrix. To estimate the rate of convergence of preconditioned iterative methods such as the Chebyshev and the conjugate gradient iterative methods we need to know the distribution of eigenvalues or, at least the bounds of extreme eigenvalues of the preconditioned matrices see [1], [2], [4], [9] [10], [14]. In the past decade the eigenvalue estimate for block incomplete factorization method has been extensively studied for symmetric positive definite matrices [3], [6], [7], [8], [11] and [12]. Recently, the authors unified some previous results and obtained some strong results even if the assumption of A is weakened.

Let B and D be symmetric matrices. Throughout the paper $B \geq D$ means that $B - D$ is positive semidefinite and $\lambda_i(A)$ denotes the ith eigenvalue of A. The eigenvalues of a matrix are ordered in a non-increasing order. Let A, X, and L be partitioned in blocks according to a given partitioning of vectors, and let $m \times m$ be the numbers of matrix blocks. The fundamental result in [5] is that

$$\lambda_i(M(\beta)) \leq \lambda_i(C^{-1}A) \leq \lambda_i(M(\sigma))$$

where $M(\alpha) = (I+\tilde{L})^{-1}+(I+\tilde{L}^T)^{-1}+(\alpha-2)(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1}$, $\tilde{L} = X^{-\frac{1}{2}}LX^{-\frac{1}{2}}$, σ and β are constant such that $\beta X + L + L^T \leq A \leq \sigma X + L + L^T$. Applying this

* This chapter is based on the paper, H. Lu and O. Axelsson, *Conditioning analysis of block incomplete factorizations and its application to elliptic equations*, Report 9504, February 1995, Department of Mathematics, University of Nijmegen, The Netherlands (submitted).

result, we obtained some simple upper bounds for the maximum eigenvalue, for example, $1/(2-\sigma)$ if $\sigma < 2$ and $2m$ if $\sigma \leq 2$ and A is positive semidefinite.

This paper continues our analysis of block incomplete factorization methods. We show that if there exist two constants α_i and σ_i such that $\alpha_i \leq \lambda_i(X^{-1}K) \leq \sigma_i$, where $K = A - L - L^T$, then

$$\lambda_{\min}(M(\alpha_i)) \leq \lambda_i(C^{-1}A) \leq \lambda_{\max}(M(\sigma_i)).$$

The result yields immediately a simple upper bound $1/(2-\sigma_i)$ for the ith eigenvalue of the preconditioned matrix $C^{-1}A$ if $\sigma_i < 2$, which result generalizes the old one for the maximum eigenvalue. Considering the relation between A and C and using the generalization of the well-known inequality that the spectral radius is bounded by the trace for symmetric positive semidefinite matrices to block form in [5], we show also that $m+1$ is another upper bound for the maximum eigenvalue if $\sigma_1 \leq 2$ and $C \leq A$. The results are used to estimate the bounds of each eigenvalue if the generalized SSOR method is used to solve an elliptic equation

$$-\frac{\partial}{\partial x}\left(a_1(x,y)\frac{\partial}{\partial x}u(x,y)\right) - \frac{\partial}{\partial y}\left(a_2(x,y)\frac{\partial}{\partial y}u(x,y)\right) = f(x,y), \quad \text{on } \Omega$$

$$u(x,y) = g(x,y) \text{ on } \Gamma = \partial\Omega,$$

where $\Omega = (0,a) \times (0,b)$, $a_1(x,y)$ and $a_2(x,y)$ are positive functions. By doing a transformation, we show how the coefficients $a_1(x,y)$ and $a_2(x,y)$ influence the bounds of eigenvalues.

8.2. Simple Upper Bounds for Every Eigenvalue

Let A be a symmetric matrix. Consider a block incomplete preconditioner $C = (X+L)X^{-1}(X+L^T)$, where X is a block diagonal and s.p.d. matrix and L is a block lower triangular matrix.

In [5] it is shown that $1/(2-\sigma)$ is a upper bound of the spectral radius of the preconditioned matrix $C^{-1}A$ if $A \leq \sigma X + L + L^T$ with $\sigma < 2$. In this section, first we generalize the result to every eigenvalue of $C^{-1}A$. To this end, we need the following basic result.

THEOREM 8.2.1. *Let A be a symmetric matrix and $C = (X+L)X^{-1}(X+L^T)$ with a s.p.d. matrix X. If there are two constants α_i and σ_i such that $\alpha_i \leq \lambda_i(X^{-1}K) \leq \sigma_i$, where $K = A - L - L^T$, then*

$$\lambda_{\min}(M(\alpha_i)) \leq \lambda_i(C^{-1}A) \leq \lambda_{\max}(M(\sigma_i)),$$

where $M(\sigma) = (I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1} + (\sigma-2)(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1}$ and $\tilde{L} = X^{-\frac{1}{2}}LX^{-\frac{1}{2}}$.

Proof. Using first a similarity transformation and then a congruence transformation, we find that the equality $\lambda_i(X^{-1}K) = \lambda_i(X^{-\frac{1}{2}}KX^{-\frac{1}{2}}) \leq \sigma_i$ is equivalent to $\lambda_i(K - \sigma_i X) \leq 0$. Furthermore, it follows from the proof of Theorem 2.1 in [5] that

$$\begin{aligned} & X^{-\frac{1}{2}}(X+L^T)C^{-1}A(X+L^T)^{-1}X^{\frac{1}{2}} \\ = \; & X^{\frac{1}{2}}(X+L)^{-1}(K-\sigma_i X)(X+L^T)^{-1}X^{\frac{1}{2}} + M(\sigma_i). \end{aligned}$$

Applying Weyl's theorem (c.f. Parlett [13] 1980, p. 192) shows that

$$\begin{aligned}\lambda_i(C^{-1}A) &= \lambda_i(X^{-\frac{1}{2}}(X+L^T)C^{-1}A(X+L^T)^{-1}X^{\frac{1}{2}}) \\ &\leq \lambda_i(X^{\frac{1}{2}}(X+L)^{-1}(K-\sigma_i X)(X+L^T)^{-1}X^{\frac{1}{2}}) + \lambda_{\max}(M(\sigma_i)),\end{aligned}$$

which implies that $\lambda_i(C^{-1}A) \leq \lambda_{\max}(M(\sigma_i))$ if $\lambda_i(X^{-1}K) \leq \sigma_i$. If $\alpha_i \leq \lambda_i(X^{-1}K)$, then $\lambda_{\min}(M(\alpha_i)) \leq \lambda_i(C^{-1}A)$ follows in a similar way. □

The following two propositions give some situations in which Theorem 8.2.1 is applicable. The first one shows a simple upper bound for every eigenvalue of the preconditioned matrix $C^{-1}A$, which generalizes some previous results for the maximum eigenvalue in [3], [5].

PROPOSITION 8.2.2. *If the conditions of* Theorem 8.2.1 *hold and* $\sigma_i < 2$, *then*

$$\lambda_i(C^{-1}A) \leq \frac{1}{2-\sigma_i}. \tag{8.2 1}$$

In particular, if $\lambda_i(X^{-1}K) < 2$, *then* $\lambda_i(C^{-1}A) \leq (2-\lambda_i(X^{-1}K))^{-1}$.

Proof. The proof is essentially the same as that of Proposition 8.2.4 in [5] □

PROPOSITION 8.2.3. *If the conditions of* Theorem 8.2.1 *hold,* $\kappa \geq \rho(M)$, *where* $M = (I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1}$, *and* $\kappa^{-\frac{1}{2}} + \sigma_i \geq 2$, *then*

$$\lambda_i(C^{-1}A) \leq (\sigma_i - 2)\kappa + 2\kappa^{\frac{1}{2}}$$

Proof. The proof is essentially the same as that of Theorem 3.3 in [5]. □

In view of the proof of Theorem 8.2.1, we find that matrix $\sigma X + L + L^T$ acts as a key to connect the symmetric matrix A and its block incomplete factorization preconditioners in estimating upper bounds of eigenvalues of the preconditioned matrix. Under the assumption $\lambda_i(K - \sigma_i X) = \lambda_i(A - \sigma_i X - L - L^T) \leq 0$, we have derived some upper bounds for every eigenvalue of the preconditioned matrix $C^{-1}A$, but such bounds can be very large if σ_i is close to 2 or they depend on the estimate of spectral radius of $M = (I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1}$. To further estimate upper bounds of eigenvalues of preconditioned matrix $C^{-1}A$ by Theorem 8.2.1, as the second part of the section, we consider the condition

$$\lambda_i(\beta_i C - \alpha_i X - L - L^T) \leq 0 \tag{8.2.2}$$

with $A \leq \sigma X + L + L^T$. The following lemma shows the relation between the parameters β_i and α_i if (8.2.2) holds.

LEMMA 8.2.4. *Let* $C = (X+L)X^{-1}(X+L^T)$, *where* X *is a s.p.d. matrix.*
1. If $\lambda_i(\beta_i C - \alpha_i X - L - L^T) \leq 0$ *with* $\alpha_i < 2$, *then* $\beta_i < 1/(2-\alpha_i)$.
2. If $\beta C \leq \alpha X + L + L^T$, *then* $\beta \leq \alpha$.

Proof. Let $A = \beta_i C$. Proposition 8.2.2 shows 1.

Let $X = \text{blockdiag}(X_1, \ldots, X_m)$. Under assumption $\beta C \leq \alpha X + L + L^T$, matrix $B = \alpha X + L + L^T - \beta C \geq 0$. Since L is a strictly lower block triangular matrix, it is readily seen that $B = \begin{pmatrix} (\alpha-\beta)X_1 & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$. Hence, $(\alpha-\beta)X_1 \geq 0$, which implies $\alpha \geq \beta$. □

COROLLARY 8.2.5. *Let* A *be a symmetric matrix and* $C = (X+L)X^{-1}(X+L^T)$, *where* X *is s.p.d. and block diagonal. If*

$$\lambda_i(\beta_i C - \alpha_i X - L - L^T) \leq 0, \quad \alpha_i < 2 \tag{8.2.3}$$

and $A \le \sigma X + L + L^T$, *then*

$$\lambda_i(C^{-1}A) \le \begin{cases} \dfrac{1}{2-\sigma}, & \text{if } \sigma \le 2 - \dfrac{2-\alpha_i}{1+\sqrt{1-\beta_i(2-\alpha_i)}} \\ \dfrac{\left(1+\sqrt{1-\beta_i(2-\alpha_i)}\right)^2(\sigma-\alpha_i)}{(2-\alpha_i)^2} + \beta_i, & \text{otherwise.} \end{cases}$$

Proof. Under the assumptions of the corollary, (8.2.3) shows for any $x > 0$

$$\lambda_i(xA + \beta_i C - (x\sigma + \alpha_i)X - (1+x)(L+L^T)) \le 0,$$

i.e.,

$$\lambda_i\left(\frac{x}{1+x}A + \frac{\beta_i}{1+x}C - \frac{x\sigma+\alpha_i}{1+x}X - (L+L^T)\right) \le 0.$$

Let $B = \frac{x}{1+x}A + \frac{\beta_i}{1+x}C$. If we choose x such that

$$\frac{x\sigma+\alpha_i}{1+x} < 2, \tag{8.2.4}$$

then $\lambda_i(B - \frac{x\sigma+\alpha_i}{1+x}X - L - L^T) \le 0$ and Proposition 8.2.2 shows that

$$\lambda_i(C^{-1}B) = \frac{x}{1+x}\lambda_i(C^{-1}A) + \frac{\beta_i}{1+x} \le \frac{1}{2 - \dfrac{x\sigma+\alpha_i}{1+x}}.$$

Hence,

$$\lambda_i(C^{-1}A) \le \frac{1+x}{x}\left(\frac{1}{2-\dfrac{x\sigma+\alpha_i}{1+x}} - \frac{\beta_i}{1+x}\right) \equiv g(x).$$

Let $y = \frac{1}{1+x}$, which implies that $0 < y < 1$. A simple computation shows that

$$f(y) = g(x) = \left(\frac{1}{2-\alpha_i} - \beta_i\right)\frac{1}{1-y} + \frac{\sigma-\alpha_i}{2-\alpha_i}\frac{1}{2-\alpha_i+(\sigma-\alpha_i)y} + \beta_i.$$

If $\sigma > 2 - \frac{2-\alpha_i}{1+\sqrt{1-\beta_i(2-\alpha_i)}}$, we have

$$\inf_{0<y<1} f(y) = f(y_0) = \frac{\left(1+\sqrt{1-\beta_i(2-\alpha_i)}\right)^2(\sigma-\alpha_i)}{(2-\alpha)^2} + \beta_i,$$

where

$$0 < y_0 = \frac{\sigma-\alpha_i-(2-\alpha_i)\sqrt{1-\beta_i(2-\alpha_i)}}{(\sigma-\alpha_i)(1+\sqrt{1-\beta_i(2-\alpha_i)})} < 1.$$

(8.2.4) holds for $x_0 > 0$ such that $y_0 = \frac{1}{1+x_0}$

Otherwise,

$$\inf_{0<y<1} f(y) = f(0) = \lim_{x\to+\infty} g(x) = \frac{1}{2-\sigma}$$

In this case $\sigma \le 2 - \frac{2-\alpha_i}{1+\sqrt{1-\beta_i(2-\alpha_i)}} < 2$, (8.2.4) holds for x sufficiently large. □

In general, it is not easy to determine a parameter pair (α_i, β_i) to minimize the bound of $\lambda_i(C^{-1}A)$ given by Corollary 8.2.5. We can use Corollary 8.2.5 as follows: fix β_i (α_i) and then estimate α_i (β_i). In particular, one can choose β_i to be negative. For example, choosing $\alpha_i = 1$ shows the following proposition:

PROPOSITION 8.2.6. *Let A be a symmetric matrix and $C = (X+L)X^{-1}(X+L^T)$, where X is block diagonal and s.p.d.. If $\lambda_{n-i+1}(\gamma_i C + X + L + L^T) \geq 0$ and $A \leq \sigma X + L + L^T$, then*

$$\lambda_i(C^{-1}A) \leq \begin{cases} (1+\sqrt{1+\gamma_i})^2(\sigma-1) - \gamma_i, & \text{if } \sigma > 2 - \frac{\sigma-1}{1+\sqrt{1+\gamma_i}}, \\ \dfrac{1}{2-\sigma} & \text{otherwise.} \end{cases}$$

Proof. Let $\alpha_i = 1$ and $\beta_i = -\gamma_i$. Condition (8.2.3) is equivalent to condition $\lambda_{n-i+1}(\gamma_i C + X + L + L^T) \geq 0$. Applying Corollary 8.2.5 shows our results. □

If $\sigma \leq 2$, Proposition 8.2.6 shows in particular that

$$\lambda_i(C^{-1}A) \leq 2(1+\sqrt{1+\gamma_i}).$$

8.3. Alternative Upper Bounds

Let $X = \text{Blockdiag}(X_1, X_2, \ldots, X_m)$ and $A \geq 0$. If $A \leq \sigma X + L + L^T$ with $1 \leq \sigma \leq 2$, we showed in [5] that $2(\sigma-1)m + 2 - \sigma$ ($\leq 2m$) is another upper bound of the maximum eigenvalue of the preconditioned matrix $C^{-1}A$. Under some strong assumptions that A is a Stieltjes matrix, $L = L_A$ and other additional conditions which is stronger than $\sigma < 2$, this bound can be reduced to $m+1$ (see [11]). In this section, we extend this kind of upper bounds to symmetric matrices. Furthermore, if the result is used for a symmetric matrix A such that $0 \leq A \leq \sigma X + L + L^T$ with $1 \leq \sigma \leq 2$, we obtain an upper bound smaller than our previous bound $2(\sigma-1)m + 2 - \sigma$. In particular, if $C \leq A$, we have $\lambda_{\max}(C^{-1}A) \leq m+1$. To finish the task, we recall the generalization of the well known equality $\rho(A) \leq \text{tr}(A)$ for a symmetric matrix $A \geq 0$ see [5], for instance.

LEMMA 8.3.1. *Let $A = (A_{ij})$ be an $m \times m$ block symmetric positive semidefinite matrix. Then*

$$\rho(A) \leq \sum_{i=1}^{m} \rho(A_{ii}). \tag{8.3.1}$$

LEMMA 8.3.2. *Let $C = (X+L)X^{-1}(X+L^T)$. If $\beta C \leq \alpha X + L + L^T$ with $\alpha < 2$, where α and β are constants, then matrix $(I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1} - \beta I$ is positive semidefinite.*

Proof. Let $A = \beta C$ and $\sigma_i = \alpha$. It follows from Theorem 8.2.1 that $\beta I \leq (I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1} + (\alpha-2)(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1}$. Hence, $(I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1} - \beta I \geq (2-\alpha)(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1}$, which implies our corollary. □

Now we prove our main result of this section.

THEOREM 8.3.3. *Let A be a symmetric matrix and $C = (X+L)X^{-1}(X+L^T)$, where $X = \text{blockdiag}(X_1, X_2, \ldots, X_m)$ is s.p.d.. If $\beta C \leq \alpha X + L + L^T$ and $A \leq \sigma X + L + L^T$, where $\alpha < 2$, σ and β are constants, then*

$$\lambda_{\max}(C^{-1}A) \leq \begin{cases} \dfrac{(\sigma-\alpha)(2-\beta)}{2-\alpha}m + \beta & \text{if } \sigma > 2. \\ (\sigma-1)(2-\beta)(m-1) + \sigma & \text{if } 1 < \sigma \leq 2 \\ 1 & \text{if } \sigma \leq 1. \end{cases}$$

Proof. It follows from the proof of Corollary 8.2.5 that for any $x > 0$ we have

$$\frac{x}{1+x}A + \frac{\beta}{1+x}C - \frac{x\sigma+\alpha}{1+x}X - (L+L^T) \leq 0.$$

Let $B = \frac{x}{1+x}A + \frac{\beta}{1+x}C$. It follows from Theorem 8.2.1 that

$$\begin{aligned}
&\lambda_{\max}(C^{-1}B) = \frac{x}{1+x}\lambda_{\max}(C^{-1}A) + \frac{\beta}{1+x} \\
\leq\ &\lambda_{\max}\left((I+\tilde{L})^{-1} + (I+\tilde{L})^{-1} + \left(\frac{x\sigma+\alpha}{1+x} - 2\right)(I+\tilde{L})^{-1}(I+\tilde{L})^{-1}\right).
\end{aligned}$$

If $\sigma > 2$, choosing $x_0 = (2-\alpha)/(\sigma-2)$ and applying Lemmas 8.3.2 and 8.3.1 shows that

$$\begin{aligned}
\lambda_{\max}(C^{-1}A) &\leq \frac{x_0+1}{x_0}\lambda_{\max}\left((I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1}\right) - \frac{\beta}{x_0} \\
&= \frac{x_0+1}{x_0}\lambda_{\max}\left(I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1} - \beta I\right) + \beta \\
&\leq \frac{(\sigma-\alpha)(2-\beta)}{2-\alpha}m + \beta
\end{aligned}$$

Let $\tilde{M} = (I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1}$. Since

$$\begin{aligned}
(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1} &= (I - ((I+\tilde{L})^{-1}\tilde{L})(I - \tilde{L}^T(I+\tilde{L}^T)^{-1}) \\
&= (I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1} - I + (I+\tilde{L})^{-1}\tilde{L}\tilde{L}^T(I+\tilde{L}^T)^{-1},
\end{aligned}$$

if $1 < \sigma \leq 2$, a computation shows that

$$\begin{aligned}
&\frac{x}{1+x}\lambda_{\max}(C^{-1}A) + \frac{\beta}{1+x} \\
\leq\ &\lambda_{\max}\left((I+\tilde{L})^{-1} + (I+\tilde{L})^{-1} + \left(\frac{x\sigma+\alpha}{1+x} - 2\right)(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1}\right) \\
\leq\ &\lambda_{\max}\left[\left(2 - \frac{x\sigma+\alpha}{1+x} + \left(\frac{x\sigma+\alpha}{1+x} - 1\right)\beta\right)I + \left(\frac{x\sigma+\alpha}{1+x} - 1\right)(\tilde{M} - \beta I)\right. \\
&\left. + \left(\frac{x\sigma+\alpha}{1+x} - 2\right)(I+\tilde{L})^{-1}\tilde{L}\tilde{L}^T(I+\tilde{L}^T)^{-1}\right] \\
\leq\ &\lambda_{\max}\left[\left(2 - \frac{x\sigma+\alpha}{1+x} + \left(\frac{x\sigma+\alpha}{1+x} - 1\right)\beta\right)I + \left(\frac{x\sigma+\alpha}{1+x} - 1\right)(\tilde{M} - \beta I)\right].
\end{aligned}$$

Setting $x \to +\infty$ and using Lemma 8.3.2 and (8.3.1) shows again that

$$\begin{aligned}
&\lambda_{\max}(C^{-1}A) \\
\leq\ &\lim_{x\to+\infty}\left[\frac{1+x}{x}\lambda_{\max}\left(\left(2 - \frac{x\sigma+\alpha}{1+x} + \left(\frac{x\sigma+\alpha}{1+x} - 1\right)\beta\right)I + \right.\right. \\
&\left.\left.\left(\frac{x\sigma+\alpha}{1+x} - 1\right)(\tilde{M} - \beta I)\right) + \frac{\beta}{x}\right] \\
\leq\ &\lambda_{\max}((2 - \sigma + (\sigma-1)\beta)I + (\sigma-1)(\tilde{M} - \beta I)) \\
\leq\ &2 - \sigma + (\sigma-1)\beta + (\sigma-1)(2-\beta)m \\
=\ &(\sigma-1)(2-\beta)(m-1) + \sigma.
\end{aligned}$$

If $\sigma \leq 1$, Proposition 8.2.2 shows that $\lambda_{\max}(C^{-1}A) \leq 1$. □

If $\sigma < 2$ and $A \geq 0$, it is readily seen that there are two constants $\beta \geq 0$ and $\alpha < 2$ such that $\beta C \leq \alpha X + L + L^T$. Applying Theorem 8.3.3 shows that

$$\begin{aligned}\lambda_{\max}(C^{-1}A) &\leq (\sigma-1)(2-\beta)(m-1)+\sigma \\ &\leq 2(\sigma-1)m+2-\sigma.\end{aligned}$$

Hence, the bound given by Theorem 8.3.3 is smaller than the bound $2(\sigma-1)m+2-\sigma$ given in [5]. In particular, If $C \leq A$, applying Theorem 8.3.3 shows the following result:

COROLLARY 8.3 4. *Let $C = (X + L)X^{-1}(X + L^T) \leq A$, where the block diagonal $X = \text{blockdiag}(X_1, X_2, \ldots, X_m)$. If there is a positive constant $\sigma \leq 2$ such that $A \leq \sigma X + L + L^T$, then*

$$\text{cond}(C^{-1}A) \leq (\sigma-1)m+1. \tag{8.3.2}$$

Proof. If $\sigma < 2$, since $C \leq A \leq \sigma X + L + L^T$, Theorem 8.3.3 shows that $\lambda_{\max}(C^{-1}A) \leq (\sigma-1)m+1$.

Since X and C are s.p.d., there exists a positive constant μ such that $X \leq \mu C$. Let ε be a positive number such that $2\varepsilon\mu < 1$. In the case $\sigma = 2$, the assumption $2X + (L + L^T) \geq A \geq 0$ implies that $-(L + L^T) \leq 2X$. Therefore, we can check that

$$\frac{C}{1+\varepsilon} \leq \frac{2}{1+\varepsilon}X + (L + L^T) + \frac{2\varepsilon\mu}{1+\varepsilon}C,$$

which implies that $\frac{1-2\varepsilon\mu}{1+\varepsilon}C \leq \frac{2}{1+\varepsilon}X + L + L^T$. It follows from Theorem 8.3.3 that

$$\lambda_{\max}(C^{-1}A) \leq (2 - \frac{1-2\varepsilon\mu}{1+\varepsilon})(m-1)+2.$$

Setting $\varepsilon \to 0$ shows (8.3.2) for $\sigma = 2$. □

8.4. Application to Elliptic Equations

As an application of our results in the previous sections, we consider an elliptic equation in two dimensions

$$-\frac{\partial}{\partial x}\left(a_1(x,y)\frac{\partial}{\partial x}u\right) - \frac{\partial}{\partial y}\left(a_2(x,y)\frac{\partial}{\partial y}u\right) = f(x,y), \quad \text{on } \Omega \tag{8.4.1}$$

$$u = g(x,y) \text{ on } \Gamma = \partial\Omega,$$

where $\Omega = (0,a)\times(0,b)$, $a_1(x,y)$ and $a_2(x,y)$ are positive. Discretizing the equation by central difference schemes with a grid of mesh size h and δ and ordering the mesh lexicographically yields the following linear system

$$A\mathbf{x} = \mathbf{b},$$

where A is a block tridiagonal matrix of the form

$$A = \text{blockdiag}(A_{i,i-1}, A_{ii}, A_{i,i+1}),$$

$$A_{ii} = \text{tridiag}(a^{(i)}_{k,k-1}, a^{(i)}_{kk}, a^{(i)}_{k,k+1}), \quad A_{i,i-1} = A_{i,i+1} = \text{diag}(b^{(i)}_1, b^{(i)}_2, \ldots, b^{(i)}_n)$$

$$i = 1,2,\ldots,m, \quad m = ah^{-1} - 1,$$

$$a^{(i)}_{kk} = a_1((i+\tfrac{1}{2})h, k\delta) + a_1((i-\tfrac{1}{2})h, k\delta) + a_2(ih, (k+\tfrac{1}{2})\delta) + a_2(ih, (k-\tfrac{1}{2})\delta)$$

$$a^{(i)}_{k,k-1} = a^{(i)}_{k-1,k} = -a_2(ih, (k-\tfrac{1}{2})\delta), \quad b^{(i)}_k = -a_1((i-\tfrac{1}{2})h, k\delta),$$

$$k = 1,2,\ldots,n, \quad n = b\delta^{-1} - 1.$$

Assume first that $a_1(x,y)$ is non-increasing for x. We compute a generalized SSOR preconditioner $C=(D-L)D^{-1}(D-L^T)$ for A as follows:

$$D_1 = A_{11},$$
$$D_i = A_{ii} - A_{i,i-1}X_{i-1}A_{i-1,i} + D_i', \quad i=2,\ldots,m,$$

where X_i, $i \geq 1$ is a sparse approximation to D_i^{-1} and D_i' is a diagonal matrix determined by

$$D_i'\mathbf{e} = A_{i,i-1}(X_{i-1} - D_{i-1}^{-1})A_{i-1,i}\mathbf{e},$$

where $\mathbf{e}=(1,1,\ldots,1)^T$. Hence, we have

$$D_1\mathbf{e} = A_{11}\mathbf{e},$$
$$D_i\mathbf{e} = (A_{ii} - A_{i,i-1}D_{i-1}^{-1}A_{i-1,i})\mathbf{e}. \tag{8.4.2}$$

Let $\triangle_i = \mathrm{diag}(a_1((i+1/2)h,\delta),\ldots,a_1((i+1/2)h,n\delta))$. It is readily seen that

$$A_{ii}\mathbf{e} \geq 2\triangle_i\mathbf{e}. \tag{8.4.3}$$

Using (8.4.2), Theorem 8.3.3 and induction, we can show that

$$D_i\mathbf{e} \geq \frac{i+1}{i}\triangle_i\mathbf{e}. \tag{8.4.4}$$

To further analyze our SSOR preconditioner, we need the following lemma.

LEMMA 8.4.1. *Let $\theta(x,y)\in\Omega$ and $m_1(x)$ and $m_2(x)$ be two functions, where $m_2(x)\geq m_1(x)$. If for any x_1 and x_2 there exist η and ξ such that $x_1\leq\eta,\xi\leq x_2$ and*

$$m_1(\eta)(x_2-x_1) \leq \theta(x_2,y)-\theta(x_1,y) \leq m_2(\xi)(x_2-x_1) \tag{8.4.5}$$

uniformly in y for any $x_1\leq x_2$ in Ω, then

$$\int_{x_1}^{x_2} m_1(x)dx \leq \theta(x_2,y)-\theta(x_1,y) \leq \int_{x_1}^{x_2} m_2(x)dx. \tag{8.4.6}$$

Proof. Under the assumptions of the lemma, we have for any $x_1=z_1<z_2<\cdots<z_m=x_2$ that

$$\theta(x_2,y)-\theta(x_1,y) = \sum_1^m(\theta(z_{i+1},y)-\theta(z_i,y)) \leq \sum_1^m m_2(\xi_i)(z_{i+1}-z_i),$$

where $z_i\leq\xi_i\leq z_{i+1}$ Setting $m\to\infty$ and letting $z_{i+1}-z_i\to 0$ shows the second inequality. The first one follows in a similar way. □

Assume that there exists a bounded integrable function $\tau(x)$ such that

$$\log a_1(x_1,y) - \log a_1(x_2,y) \leq \tau(\xi)(x_2-x_1) \tag{8.4.7}$$

for any $x_2\geq x_1$ in Ω, where $x_1\leq\xi\leq x_2$. Let $\mu_i = \exp\left(\int_{(i-1/2)h}^{(i+1/2)h}\tau(x)dx\right)$ and $\beta_i = (i+1)\frac{1+\mu_i}{i+1+\mu_i}$. A computation using (8.4.2), (8.4.4) and Lemma 8.4.1 shows that

$$\begin{aligned}
\beta_i D_i\mathbf{e} &= A_{ii}\mathbf{e} + [(\beta_i-1)A_{ii} - \beta_i A_{i,i-1}D_{i-1}^{-1}A_{i-1,i}]\mathbf{e} \\
&\geq A_{ii}\mathbf{e} + \left[(\beta_i-1)(\triangle_i+\triangle_{i-1}) - \frac{i\beta_i\triangle_{i-1}}{i+1}\right]\mathbf{e} \\
&\geq A_{ii}\mathbf{e}
\end{aligned}$$

In addition, $\beta_i D_i - A_{ii} = (\beta_i - 1)D_i + (D_i - A_{ii})$ is a Z-matrix. Therefore, $\beta_i D_i - A_{ii} \geq 0$ and

$$\rho(D_i^{-1}A_{ii}) \leq \beta_i, \quad i = 1, 2, \ldots, m.$$

Reordering β_i such that $\beta_{i+1} \geq \beta_i$, we have

$$\lambda_{in+j}(D^{-1}D_A) \leq \beta_{m-i}, \quad i = 0, 1, \ldots, m-1, \ j = 1, 2, \ldots, n.$$

If $\beta_{m-i} \leq 2$, it follows from Proposition 8.2.2 that

$$\lambda_{in+j}(C^{-1}A) \leq \frac{1}{2 - \beta_{m-i}}. \tag{8.4.8}$$

In particular, if $\beta_m < 2$, (8.4.8) gives upper bounds for the maximum eigenvalue of the preconditioned matrix $C^{-1}A$. If $\tau(x) \equiv 0$, i.e., a_1 does not depend on x, then all $\beta_i < 2$ and hence (8.4.8) becomes

$$\lambda_{in+j}(C^{-1}A) \leq \frac{m-i+2}{2}, \quad i = 0, \ldots, m-1, \ j = 1, 2, \ldots, n.$$

Consider the lower block tridiagonal matrix $T = I - D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$. We have

$$T^{-1} = (T_{ij}) = \sum_{r=0}^{m-1} (D^{-\frac{1}{2}}LD^{-\frac{1}{2}})^r,$$

where $T_{ii} = I$, $T_{ij} = D_i^{-\frac{1}{2}}\Delta_{i-1}D_{i-1}^{-1}\cdots D_{j+1}^{-1}\Delta_j D_j^{-\frac{1}{2}}$ for $i > j$ and $T_{ij} = 0$ for $i < j$. Partition $M = T^{-1}(T^T)^{-1}$ into an $m \times m$ block matrix (M_{ij}) consistently with the partitioning of A, where M stands for the same matrix as in Proposition 8.2.3. Applying (8.4.4) and Lemma 8.4.1 shows that

$$\begin{aligned}
& D_i^{-1}\Delta_i \mathbf{e} \leq \frac{i}{i+1}\mathbf{e}, \\
& D_i^{-1}\Delta_{i-1}\mathbf{e} = D_i^{-1}\Delta_i\Delta_i^{-1}\Delta_{i-1}\mathbf{e} \\
= \ & D_i^{-1}\Delta_i \operatorname{diag}\left(\frac{a_1((i-1/2)h, h)}{a_1((i+1/2)h, h)}, \cdots, \frac{a_1((i-1/2)h, nh)}{a_1((i+1/2)h, nh)}\right)\mathbf{e} \\
\leq \ & \mu_i D_i^{-1}\Delta_i \mathbf{e} \leq \frac{i\mu_i}{i+1}\mathbf{e}
\end{aligned}$$

Therefore, for $i \geq k$,

$$\begin{aligned}
& D_i^{-\frac{1}{2}}T_{ik}T_{ik}^T D_i^{\frac{1}{2}}\mathbf{e} \\
= \ & (D_i^{-1}\Delta_{i-1})\cdots(D_{k+1}^{-1}\Delta_k)(D_k^{-1}\Delta_k)\cdots(D_{i-1}^{-1}\Delta_{i-1})\mathbf{e} \\
\leq \ & \frac{k(k+1)}{i(i+1)}\prod_{j=k}^{i-1}\mu_j \mathbf{e} \\
= \ & \frac{k(k+1)}{i(i+1)}\exp\left(\int_{(k+1/2)h}^{(i+1/2)h}\tau(z)dz\right)\mathbf{e},
\end{aligned}$$

which implies that the spectral radius

$$\rho(T_{ik}T_{ik}^T) = \rho(D_i^{-\frac{1}{2}}T_{ik}T_{ik}^T D_i^{\frac{1}{2}}) \leq \frac{k(k+1)}{i(i+1)}\exp\left(\int_{(k+1/2)h}^{(i+1/2)h}\tau(z)dz\right),$$

because $D_i^{-\frac{1}{2}} T_{ik} T_{ik}^T D_i^{-\frac{1}{2}}$ is a nonnegative matrix. It follows from Lemma 8.3.1 that

$$\begin{aligned}
\rho(M) &\le \sum_{i=1}^{m} \rho(M_{ii}) = \sum_{i=1}^{m} \rho\left(\sum_{k=1}^{i} T_{ik} T_{ik}^T\right) \le \sum_{i=1}^{m}\sum_{k=1}^{i} \rho(T_{ik} T_{ik}^T) \\
&\le \sum_{i=1}^{m}\sum_{k=1}^{i} \frac{k(k+1)}{i(i+1)} \exp\left(\int_{(k+1/2)h}^{(i+1/2)h} \tau(z)dz\right) \\
&\le \sum_{i=1}^{m}\sum_{k=1}^{i} \left(\frac{k+1/2}{i+1/2}\right)^2 \exp\left(\int_{(k+1/2)h}^{(i+1/2)h} \tau(z)dz\right) \\
&\le \left(\frac{m+1}{a}\right)^2 \sum_{i=1}^{m}\sum_{k=1}^{i} \frac{a(i+1)}{m+1} \left(\frac{k+1/2}{i+1/2}\right)^2 \\
&\quad \exp\left(\int_{\frac{k+1/2}{i+1/2}\frac{a(i+1/2)}{m+1/2}}^{\frac{a(i+1/2)}{m+1/2}} \tau(z)dz\right) \frac{1}{i+1/2}\frac{a}{m+1/2} \\
&\le h^{-2} \int_0^a y \int_0^1 x^2 \exp\left(\int_{xy}^{y} \tau(z)dz\right) dxdy + C_1 h^{-1} + C_2 \equiv \kappa,
\end{aligned}$$

where C_1 and C_2 are constants independent of h. Let $\nu = \max_{0\le x\le a} \tau(x)$. It follows by a simple computation that

$$\sigma_1 = \max_{1\le i\le m} \beta_i = \max_{1\le i\le m} (i+1)\frac{1+\mu_i}{i+1+\mu_i} \le 2 + (\nu-2)h + o(h),$$

Proposition 8.2.3 yields the following upper bound for the maximum eigenvalue of preconditioned matrix:

$$\begin{aligned}
\lambda_{\max}(C^{-1}A) &\le (\sigma_m - 2)\kappa + 2\kappa^{\frac{1}{2}} \\
&\le \begin{cases} 2C^{\frac{1}{2}} h^{-1} + C_3, & \text{if } \sigma_1 \le 2, \\ [(\nu-2)C + 2C^{\frac{1}{2}}]h^{-1} + C_4, & \text{if } \sigma_1 > 2, \end{cases}
\end{aligned}$$

where $C = \int_0^a y \int_0^1 x^2 \exp\left(\int_{xy}^{y} \tau(z)dz\right) dxdy$, C_3 and C_4 are constants independent of h.

For the general case, assume that there exist two bounded integrable functions such that

$$\varphi_1(\mu)(x_2 - x_1) \le \log a_1(x_2, y) - \log a_2(x_1, y) \le \varphi_2(\xi)(x_2 - x_1) \tag{8.4.9}$$

uniformly in y for any $x_2 > x_1$ in Ω. Note that it is straightforward to show that (8.4.9) is equivalent to that the function $\log a_1(x, y)$ satisfies the Lipschitz condition

$$|\log a_1(x_2, y) - \log a_1(x_1, y)| \le L|x_2 - x_1| \tag{8.4.10}$$

uniformly in y for any $x_1, x_1 \in \Omega$. where L is a positive constant. This implies that we can allow that the function $a_1(x, y)$ has jumps in y-direction. Defining $\chi(x) = \exp\left(\int_0^x \varphi_2(t)dt\right)$ and making a transformation $w = \int_0^x 1/\chi(t)dt$, we obtain the equation

$$-\frac{\partial}{\partial w}\left(b_1 \frac{\partial}{\partial w} u\right) - \frac{\partial}{\partial y}\left(b_2 \frac{\partial}{\partial y} u\right) = \tilde{f} \quad \text{on } \tilde{\Omega},$$

$$u = g \quad \text{on } \tilde{\Gamma} = \partial\tilde{\Omega},$$

where $b_1(w,y) = a_1(x,y)/\chi(x)$, $b_2(w,y) = \chi(x)a_2(x,y)$, $\tilde{f}(w,y) = \chi(x)f(x,y)$ and $\tilde{\Omega} = (0,\tilde{a}) \times (0,b)$ with $\tilde{a} = \int_0^a 1/\chi(x)dx$.

We then discretize the equation by a central difference scheme and construct a preconditioner of the same form as previously.

Note that for any $w_2 = \int_0^{x_2} 1/\chi(t)dt$ and $w_1 = \int_0^{x_1} 1/\chi(t)dt$ we have $x_2 > x_1$ if $w_2 > w_1$ simply because $\chi(x) > 0$. Hence, for any $w_2 > w_1$, using Lemma 8.4.1, we have

$$\begin{aligned} b_1(w_2,y) - b_1(w_1,y) &= \frac{a_1(x_2,y)}{\chi(x_2)} - \frac{a_1(x_1,y)}{\chi(x_1)} \\ &\le \frac{a_1(x_1,y)}{\chi(x_2)}\left(\frac{a_1(x_2,y)}{a_1(x_1,y)} - \frac{\chi(x_2)}{\chi(x_1)}\right) \\ &\quad \frac{a_1(x_1,y)}{\chi(x_2)}\left(\frac{a_1(x_2,y)}{a_1(x_1,y)} - \exp\left(\int_{x_1}^{x_2}\varphi_2(x)dx\right)\right) \\ &\le 0, \end{aligned}$$

which implies that $b_1(w,y)$ is non-increasing for w.

On the other hand, for $w_2 > w_1$ a computation shows that

$$\begin{aligned} & \log b_1(w_1,y) - \log b_1(w_2,y) \\ = \;& \log a_1(x_1,y) - \log a_1(x_2,y) + \log\chi(x_2) - \log\chi(x_1) \\ \le \;& -\int_{x_1}^{x_2}\varphi_1(x)dx + \int_{x_1}^{x_2}\varphi_2(x)dx \\ = \;& \int_{x_1}^{x_2}(\varphi_2(x) - \varphi_1(x))dx \\ \le \;& \int_{w_1}^{w_2}(\varphi_2(x) - \varphi_1(x))\exp\left(\int_0^x \varphi_2(t)dt\right)dw \end{aligned}$$

Let $\psi(w) = (\varphi_2(x) - \varphi_1(x))\exp\left(\int_0^x \varphi_2(t)dt\right)$, where $w_i = \int_0^{x_i} 1/\chi(t)dt$. Our previous analysis shows that

$$\lambda_{\max}(C^{-1}A) \le \begin{cases} 2D^{\frac{1}{2}}h^{-1} + D_1, & \text{if } \gamma_1 \le 2, \\ [(\mu-2)D + 2D^{\frac{1}{2}}]h^{-1} + D_2, & \text{if } \gamma_1 > 2, \end{cases}$$

where $D = \int_0^{\tilde{a}} y \int_0^1 x^2 \exp\left(\int_{xy}^{y}\psi(t)dt\right)dxdy$, $\mu = \max_{0\le w\le\tilde{a}}\psi(w)$, D_1 and D_2 are constants independent of h and

$$\gamma_1 = \max_{0\le i\le m}(i+1)\frac{1 + \exp\left(\int_{(i-1/2)h}^{(i+1/2)h}\psi(x)dx\right)}{i+1+\exp\left(\int_{(i-1/2)h}^{(i+1/2)h}\psi(x)dx\right)}.$$

The above shows that the upper bound $O(h^{-1})$ of the maximum eigenvalue of $C^{-1}A$ holds also for problems (8.4.1) with variable coefficients, if $\log a_1(x,y)$ satisfies the Lipschitz condition (8.4.10) for x. Because $A - C$ is a Z-matrix and $(A-C)\mathrm{e} = 0$ for the lower eigenvalue bound we have $\lambda_{\min}(C^{-1}A) \ge 1$. Therefore, all upper bounds for the maximum eigenvalue in this section give also upper bounds of the condition number for the preconditioned matrix $C^{-1}A$.

8.5. Conclusions

The application of eigenvalue estimates yields an $O(h^{-1})$ upper bound for the condition number of the preconditioned matrix if the modified block factorization method is used for elliptic equations (8.4.1) with variable coefficients under the assumption that $\log a_1(x, y)$ satisfies the Lipschitz condition (8.4.10) for x. If $\log a_2(x, y)$ satisfies a Lipschitz condition for y, we can reorder the mesh and use the same approach to the new discrete linear system that yields an $O(\delta^{-1})$ upper bound for the condition number of the new preconditioned matrix. In particular, this is an efficient way to solve equation (8.4.1) if the coefficient $a_1(x, y)$ does not satisfy the conditions required. Finally, it turns out that the results and the approach can also be used to equation (8.4.1) subject to certain forms of mixed boundary conditions.

References

[1] O. AXELSSON, *A class of iterative methods for finite element equations*, Comput. Methods Appl. Mech. Engrg., 9 (1976), pp. 123–137.

[2] ———, *Solution of linear systems of equations: iterative methods*, in Sparse Matrix Techniques, Lecture Notes in Mathematics 572, V. A. Barker, ed., Springer Verlag, Berlin, Heidelberg, New York, 1977, pp. 1–50.

[3] ———, *Bounds of eigenvalues of preconditioned matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 847–862.

[4] O. AXELSSON AND G. LINDSKOG, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 48 (1986), pp. 499–523.

[5] O. AXELSSON AND H. LU, *On eigenvalue estimates for block incomplete factorization methods*, SIAM J. Matrix Anal. Appl., 16 (1995), p. (to appear).

[6] R. BEAUWENS, *Upper eigenvalue bounds for pencils of matrices*, Linear Algebra Appl., 62 (1984), pp. 87 104.

[7] ———, *Approximate factorizations with S/P consistently ordered M-factors*, BIT, 29 (1989), pp. 658–681.

[8] ———, *Approximate factorizations with modified S/P consistently ordered M-factors*, Numer. Linear Algebra Appl., 1 (1994), pp. 3–17.

[9] A. GREENBAUM, *Comparison of splittings used with the conjugate gradient algorithm*, Numer. Math., 33 (1979), pp. 181–194.

[10] A. JENNINGS, *Influence of the eigenvalues spectrum of the convergence rate of the conjugate gradient method*, IMA Journal of Numerical Analysis, 20 (1977), pp. 61–72

[11] M. M. MAGOLU, *Analytical bounds for block approximate factorization methods*, Linear Algebra Appl, 179 (1993), pp. 33–57.

[12] M. M. MAGOLU AND Y. NOTAY, *On the conditioning analysis of block approximate factorization methods*, Linear Algebra Appl., 154-156 (1991), pp. 583–599.

[13] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, N.J., 1980.

[14] A. VAN DER SLUIS AND H. A. VAN DER VORST, *The rate of convergence conjugate gradients*, Numer. Math., 48 (1986), pp. 543–560.

Samenvatting

Het proefschrift bestaat uit twee delen, in totaal bestaande uit acht hoofdstukken. Het kwam tot stand in een periode van twee en een half jaar van eind 1992 tot begin 1995 waarin ik werkzaam was als OIO onder leiding van Professor O. Axelsson.

Het eerste deel, bestaande uit 4 hoofdstukken, handelt over de existentie en uniciteit van een zwakke oplossing en eindige elementen methoden voor voorwaarts-achterwaartse warmtevergelijkingen en een consistentie grens op eindige differentie schema's van positief type voor convectie-diffusie problemen. Het tweede deel, ook 4 hoofdstukken, houdt zich voornamelijk bezig met het gebruik van de numerieke radius om de convergentiesnelheid van iteratieve methoden voor niet-symmetrische lineaire systemen te analyseren, generalisaties van diagonale compensatie, schattingen van conditiegetallen en eigenwaarden voor blok incomplete factorisatie methoden met toepassingen op elliptische vergelijkingen. Er volgt nu een korte samenvatting van ieder van de delen.

Deel 1

In de literatuur zijn al veel eindige elementen methoden voor de warmtevergelijking geïntroduceerd en geanalyseerd. Gewoonlijk past men eerst een Galerkin methode toe in de ruimtevariabelen waardoor de vergelijkingen reduceren tot een stelsel gewone differentiaalvergelijkingen. Vervolgens past men een geschikte methode toe om de gewone differentiaalvergelijkingen te integreren. Helaas zijn er warmtevergelijkingen die niet in dit schema passen, bijvoorbeeld de voorwaarts-achterwaartse warmtevergelijkingen.

Voor een beter begrip van de voorwaarts-achterwaartse warmtevergelijkingen bewijzen we eerst de existentie en uniciteit van een zwakke oplossing van de vergelijkingen gegeven passende randvoorwaarden in een zekere Hilbertruimte en we laten zien dat het probleem in een zekere zin welgesteld is. Het bewijs is gebaseerd op een variant van het Lax–Milgram lemma. Na deze analyse beschouwen we eindige elementen methoden, met name ruimte-tijd methoden, voor de voorwaarts-achterwaartse warmtevergelijkingen. We bekijken gelijktijdige discretisatie van ruimte- en tijdvariabelen gebruikmakend van continue eindige elementen methoden. De resulterende lineaire stelsels zijn positief definiet. Deze methoden hebben duidelijke voordelen over bestaande methoden voor deze problemen. Om een brede klasse van voorwaarts-achterwaartse warmtevergelijkingen te kunnen oplossen gebruiken we variabele-transformaties zodat de nieuwe vergelijkingen kunnen worden opgelost met onze ruimte-tijd methoden. We leiden voorwaarden af voor de toepasbaarheid van de transformaties en laten zien hoe de transformaties kunnen worden geconstrueerd.

We bepalen tevens een consistentie grens op eindige differentie schema's van positief type voor convectie-diffusie problemen.

Deel 2

In dit deel analyseren we de standaard iteratieve methoden en de SOR methode voor quasi-Hermitische positief definiete matrices gebruikmakend van de numerieke radius, we generaliseren het gebruik van diagonale compensatie van symmetrische positief definiete matrices naar positief definiete matrices en we geven schattingen van eigenwaarden voor blok incomplete factorisatie methoden en compenserende preconditioneringen.

Gewoonlijk wordt de spectraalstraal gebruikt om de convergentiesnelheid van iteratieve methoden te analyseren. Echter, voor een niet-symmetrische iteratiematrix, geef dit allen informatie over het asymptotische gedrag. We laten eerst zien dat de numerieke radius een betere maat is voor de convergentiesnelheid voor de initiële iteraties. In de analyse van de successieve overrelaxatie methode voor quasi-Hermitische positief definiete matrices laten we zien dat een cruciale parameter afhangt van de numerieke radius van het (blok) benedendriehoeksdeel van de matrix. In het geval van blok incomplete factorisatie geven we een bovengrens voor de grootste eigenwaarde van de gepreconditioneerde matrix met behulp van de numerieke radius.

We introduceren matrixcompensatie om de positiviteit van matrices te behouden en laten zien hoe diagonale compensatie gebruikt kan worden voor niet-symmetrische matrices. We geven schattingen van conditiegetallen van gepreconditioneerde matrices in het geval van diagonale compensatie bij het construeren van preconditioneringsmatrices voor symmetrische positief definiete matrices.

Voor symmetrische matrices geven we schattingen van bovengrenzen en verdeling van de eigenwaarden van blok incomplete factorisaties. We laten zien hoe de diagonaal en driehoeksdelen van de matrices de eigenwaarden beïnvloeden. Zelfs als de voorwaarden op de matrices worden afgezwakt hebben we scherpe resultaten. De resultaten die we hier presenteren unificeren enkele van de eerdere resultaten betreffende bovengrenzen voor de eigenwaarden van incomplete preconditioneringen. Ook generaliseren we de bekende ongelijkheid dat de spectraalstraal begrensd wordt door het spoor voor symmetrische positief semidefiniete matrices naar blokvorm. De resultaten worden gebruikt om elliptische problemen op te lossen. We laten zien hoe de coëfficiënten in de elliptische vergelijkingen de grenzen op de eigenwaarden beïnvloeden. De analyse van de eigenwaardeschattingen levert een $O(h^{-1})$ bovengrens voor het conditiegetal van de gepreconditioneerde matrix als we de gemodificeerde blok incomplete factorisatie gebruiken voor elliptische vergelijkingen met variabele coëffiënten.

Curriculum Vitae

Ik ben geboren op 1 november 1961 in Shaanxi, Volksrepubliek China. In 1979 ben ik wiskunde gaan studeren aan de universiteit van Lanzhou. Na vier jaar studie kreeg ik in 1983 een aanstelling als assistent docent bij de afdeling Wiskunde van de universiteit van Xi'an Jiaotong. Van 1985 tot 1988 studeerde ik voor mijn doctoraal wiskunde naast mijn werkzaamheden als docent, begin 1988 afgesloten met een doctoraaldiploma. Aan het eind van dat jaar werd ik gepromoveerd tot docent wiskunde aan de universiteit van Xi'an Jiaotong. Op 1 november 1992 kreeg ik in Nijmegen een aanstelling als OIO onder leiding van Professor Axelsson.